Faculty of Sciences
Department of Biochemistry & Microbiology
Laboratory for Protein Biochemistry and Biomolecular Engineering (L-ProBE)

Academic year 2014–2015

Novel methods for C-terminal sequence analysis in the proteome era

Pablo Moerman

Promotors:
Prof. dr. B. Devreese
Dr. B. Samyn

Thesis submitted in fulfillment of the requirements
for the degree of Doctor of Science: Biochemistry

Promotors:

Prof. dr. Bart Devreese
Laboratory for Protein Biochemistry and Biomolecular Engineering (L-ProBE)
Department of Biochemistry and Microbiology
Faculty of Science, Ghent University, Belgium

Dr. Bart Samyn
Laboratory for Protein Biochemistry and Biomolecular Engineering (L-ProBE)
Department of Biochemistry and Microbiology
Faculty of Science, Ghent University, Belgium

Examination Committee:

Prof. dr. Savvas Savvides (Chairman)
Department of Biochemistry and Microbiology
Faculty of Science, Ghent University, Belgium

Prof. dr. Annemieke Madder
Department of Organic and Macromolecular Chemistry
Faculty of Science, Ghent University, Belgium

Prof. dr. Sebastien Carpentier
Department of Biosystems
Faculty of Bioscience Engineering
KU Leuven, Belgium

Dr. Kris Morreel
VIB-department of Plant Systems Biology
Department of Biotechnology and Bio-informatics
Faculty of Science, Ghent University, Belgium

Dr. Kjell Sergeant
Department Environment and Agro-biotechnologies
Centre de Recherche Public, Gabriel Lippmann, Luxembourg

Prof. dr. em. Jos Van Beeumen
Department of Biochemistry and Microbiology
Faculty of Science, Ghent University, Belgium

This work was supported by a Ph.D. grant of the Institute for the promotion of Innovation through Science and Technology in Flanders (I.W.T.-Vlaanderen).

Dutch translation of the title:
Nieuwe methodes voor C-terminale sequentie analyse in het proteoom tijdperk.

Printed by University press, Zelzate, Begium
www.universitypress.be

ISBN

Dankwoord

Bij het beëindigen van dit proefschrift gaat mijn dank uit naar een aantal personen, die een belangrijke rol hebben gespeeld in de realisatie ervan.

In de eerste plaats wens ik mijn promotor Prof. dr. B. Devreese te bedanken voor de kans die u mij geboden heeft om dit onderzoek te starten in het laboratorium voor Eiwitbiochemie en Biomoleculaire Engineering (L-ProBE). Niet enkel het voorzien van de nodige faciliteiten, maar tevens uw begeleiding hebben geholpen bij het voltooien van dit project.

Daarnaast wens ik bijzondere dankbaarheid te getuigen aan Dr. B. Samyn. Uw project, inzicht en gerichte voorbereiding hebben mij klaargestoomd voor een succesvolle IWT-verdediging. Dankzij deze inspanningen kreeg ik de kans te starten aan dit onderzoek. Onze ietwat verschillende interpretatie van het begrip deadline heeft zo nu en dan wel voor wat stress gezorgd. Bedankt om me desondanks te blijven ondersteunen met uw denkpistes, analyses en commentaren.

Verder wens ik de leden van de lees- en examencommissie te bedanken voor de grondige lezing van het manuscript en de constructieve commentaren die hebben bijgedragen tot deze finale versie. In het bijzonder wens ik Prof. dr. em. J. Van Beeumen te bedanken voor zijn uitvoerige bijdrage als examencommissaris.

Uiteraard dank ik ook de collega's van het lab. De meeste zijn er ondertussen al vandoor, slechts een aantal getrouwen zijn achtergebleven. Gonzy, onze road trip in California zal ik niet snel vergeten en is wat mij betreft voor herhaling vatbaar. Isaak en Griet, ik vrees dat jullie mij altijd zullen blijven associëren met de geur van broodjes Noordzee salade en droge worstjes, mijn excuses daarvoor. Ik heb goede herinneringen aan ons bureautje samen, de koude winters op het 6^e, daarna klein en gezellig op het 2^e. Alexandra, Jonathan, Ester, Bert, Leander, Silke, jullie zijn in de loop der jaren meer geworden dan gewoon collega's. Bedankt voor de aangename jaren! Ik zal in de toekomst wat meer tijd vrijmaken voor jullie. Verder wens ik ook de overige collega's te bedanken voor de aangename en stimulerende werksfeer. Hierbij denk ik niet enkel aan de medewerkers van mijn eigen groep, maar ook aan de collega's uit de groep van Prof. dr. S. Savvides en een aantal microbiologen. Ik houd vooral goede

herinneringen aan jullie over.

Dan rest er nog een omvangrijke groep vrienden, (ex-)huisgenoten en familieleden. De meesten onder jullie heb ik de laatste maanden serieus verwaarloosd, ik probeer dit de komende maanden zeker goed te maken. Lore, bedankt om zoveel geduld te hebben.

Tot slot mag ik zeker mijn huidige collega's van Centexbel niet vergeten te bedanken. Twee collega's in het bijzonder verdienen vermelding. David, bedankt om mij op cruciale momenten bij te staan met je gedegen organische chemiekennis. Eddy, het is dankzij de flexibiliteit die u mij het afgelopen jaar gegeven heeft dat ik dit heb kunnen afwerken. Van harte bedankt daarvoor.

Pablo
Gent, 2014

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2D-PAGE	Two-Dimensional Polyacrylamide Gel Electrophoresis
ACN	Acetonitrile
AQUA	Absolute Quantification
ATZ	Anilinothiazolinone
BLOSUM	BLOcks SUbstitution Matrix
BNPS	2-(2-Nitrophenylsulfenyl)-3‘-methyl-3-Bromoindolenine
CAD	Collisionally Activated Dissociation
CHAPS	3-[(cholamidopropyl) dimethylammonio]- 1-propanesulfonate
CID	Collision Induced Dissociation
CNBr	Cyanogen Bromide
COFRADIC	Combined Fractional Diagonal Chromatography
CPase	Carboxypeptidase
CPP	Carboxypeptidase P
CPY	Carboxypeptidase Y
C-TAILS	C-Terminal Amine-based Isotope Labeling of Substrates
DCC	N,N’- Dicyclohexylcarbodiimide
DIC	N,N’-Diisopropylcarbodiimide
DICAS	Dimethyl Isotope-Coded Affinity Selection
DIEA	N,N-diisopropylamine
DIGE	Difference Gel Electrophoresis
DITC	Diisothiocyanate
DMF	Dimethylformamide
DNA	Deoxyribonucleic acid
DPP-ITC	Diphenyl phosphoroisothiocyanatidate
DTT	Dithiothreitol
ECD	Electron Capture Dissotiation
EDC	1-(3- dimethylaminopropyl)-3-ethylcarbodiimide hydrochloride
EDTA	Ethylenediaminetetraacetic acid
ETD	Electron Transfer Dissociation
ETD-LTQ	Electron Transfer Dissociation - Linear Trap Quadrupole
ESI	Electrospray Ionization
FEP	Fluorinated Ethylene Propylene
FT-ICR	Fourier Transform - Ion Cyclotron Resonance
GeLC	Gel Electrophoresis - Liquid Chromatography
GPI	Glycophosphatidylinositol
HOAt	1-Hydroxy-7-Azabenzotriazole

HPLC	High Performance Liquid Chromatography
HUPO	Human Proteome Organization
IAA	Iodoacetic acid
ICAT	Isotope Coded Affinity Tag
IEF	Iso-electric Focusing
IPG	Immobilized pH Gradient
IRMPD	Infrared Multiphoton Dissotiation
iTRAQ	Isobaric Tagging for Relative and Absolute Quantification
LB	Luria Bertani
LC	Liquid Chromatography
MALDI	Matrix Assisted Laser Desorption Ionization
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
MudPIT	Multidimensional Proteome Identification Technique
Mw	Molecular Weight
m/z	Mass-to-Charge ratio
NBS	N-Bromosuccinimide
NCBI	National Center for Biotechnology Information
N-CLAP	N-terminalomics by Chemical Labeling of the α-Amine of Proteins
NHS	N-hydroxysuccinimide
NIS	N-iodosuccinimide
ORF	Open Reading Frame
PBS	Phosphate Buffered Saline
PFF	Peptide Fragment Fingerprinting
pGAPase	Pyroglutamyl Aminopeptidase
pI	Isoelectric Point
PIC	Phenylisocyanate
PICS	Proteomic Identification of Protease Cleavage Sites
PITC	Phenylisothiocyanate
PMF	Peptide Mass Fingerprint
PP	1-(2-pyrimidyl)piperazine
ProC-TEL	Profiling Protein C-Termini by Enzymatic Labeling
PSP	Proteolytic Signature Peptide
PTC	Phenylthiocarbamyl
PTFE	Polytetrafluoroethylene
PTH	Phenylthiohydanthoin
PTM	Post Translation Modification
PVDF	Polyvinylidene Fluoride
Q	Quadrupole
Qcyclase	Glutamine Cyclotransferase
RNA	Ribonucleic acid
RP	Reversed phase
RT	Room Temperature
SCX	Strong Cation Exchange
SDS	Sodium Dodecyl Sulfate
SILAC	Stable Isotope Labeling by Amino Acids in Cell Culture

S-NHS	Sulfo-N-hydroxysuccinimide
STEP	STandard of Expressed Protein
TAILS	Terminal Amine-based Isotope Labeling of Substrates
TCEP	Tris(2-carboxyethyl)phosphine
TFA	Trifluoroacetic acid
TH	thiohydantoin
TMPP	Tris-(2,4,6-trimethoxyphenyl)Phosphonium
TNBS	2,4,6-Trinitrobenzenesulfonic Acid
TOF	Time Of Flight
TopFIND	Terminus Oriented Protein Function INferred Database
TPR	Tetratricopeptide Repeat
Tris	Tris(hydroxymethyl)aminomethane

Samenvatting

De verwezenlijkingen binnen het genoomonderzoek, met ondermeer de publicatie van de volledige sequentie van het menselijke genoom, bieden een schat aan informatie voor het biochemisch en biomedisch onderzoek. Naarmate echter de genetische kennis completer werd verschoof de wetenschappelijke aandacht meer en meer naar de biologische interpretatie van deze informatie. Het onderzoeksveld 'proteoomanalyse' omvat het grootschalig identificeren en kwantificeren van de uiteindelijke genproducten, de eiwitten, maar ook de analyse van hun modificaties. De studie van het proteoom is bijzonder uitdagend gezien de grote dynamische verschillen waarmee eiwitten tot expressie gebracht worden en de grote variabiliteit in de finale structuur van de eiwitten; 'splice'-varianten, N- en C-terminale processing, co- en post-translationele modificaties (PTMs) [1]. N- en C-terminale splitsingen in de gevormde polypeptidenketen behoren tot de meest voorkomende types van PTMs in eiwitten. Hoewel deze modificaties een grote impact kunnen hebben op de biologische activiteit van eiwitten, wordt hier nog relatief weinig aandacht aan besteed en is er nood aan de ontwikkeling van bijkomende technieken die een systematische analyse van proteolytische processing toelaten.

Het systematisch bepalen van de eiwittermini kan bovendien ook bijdragen tot een meer accurate en betrouwbare genoomannotatie. Om meer betrouwbare voorspellingen uit te voeren wordt momenteel voor de annotatie van genen gebruik gemaakt van een combinatie van verschillende algoritmes voor gen-predictie en similariteitsanalyse. In proteogenomics wordt data afkomstig uit proteoomanalyse gebruikt als extra parameter om de theoretisch voorspelde genen te kunnen bevestigen of corrigeren [2–6]. De correcte C-terminus van het finale genproduct bepalen is een van de grootste uitdagingen binnen dit onderzoeksdomein.

In de loop der jaren zijn verschillende methodes ontwikkeld om de C-terminale sequentie van een eiwit te bepalen. De oorspronkelijke benaderingen omvatten chemische sequentiemethoden, complementair aan de Edmandegradatie, maar deze waren minder gevoelig en betrouwbaar [7]. Nadien werden een aantal massaspectrometrische technieken ontwikkeld, waaronder het uitvoeren van een beperkte eiwithydrolyse [8] of de analyse van in-source decay fragmenten in een MALDI-TOF massaspectrometer uitgerust met LIFT-concept [9]. Eveneens werden een aantal technieken ontwikkeld die toelaten om het C-terminale peptide te isoleren of selectief te merken; voorbeelden hiervan zijn het gebruik van anhydrotrypsinekolommen [10] en

de merking van tryptische peptiden met ^{18}O isotopen [11]. Meer recent werden een aantal LC-MS gebaseerde technieken beschreven waaronder het gebruik van een ionenuitwisseling om N-terminaal geblokkeerde peptiden en C-terminale peptiden aan te rijken [12] alsook het gebruik van de COFRADIC technologie om specifiek N- en C-terminal peptiden te selecteren via diagonale chromatografie [13]. Geen van deze methodes is echt doorgebroken als standaardtechnologie voor C-terminale sequentiebepaling en er is dus nog steeds nood aan eenvoudige, gevoelige, en robuuste analysetechnieken.

Het laboratorium voor Eiwitbiochemie en Biomoleculaire Engineering (L-ProBE) heeft een rijke geschiedenis in het ontwikkelen van methoden voor C-terminale sequentieanalyse. In 2005 werd een C-terminale sequentieanalyse techniek ontwikkeld die gebruik maakt van carboxypeptidasen om C-terminale sequentieladders te genereren in peptidefragmenten bekomen na behandeling van eiwitten met cyanogeen bromide (CNBr) [14, 15]. Tijdens de splitsing van eiwitten met CNBr worden alle peptidebindingen C-terminaal van methionine verbroken en wordt het terminale methionine omgezet naar een homoserinelacton. Het C-terminale peptide is het enige dat een carboxyl terminus heeft en gedegradeerd kan worden door carboxypeptidasen. Het doel van dit project was deze techniek verder te verfijnen door enkele van de belangrijkste beperkingen weg te werken en zo een geautomatiseerd C-terminale sequentieanalyse platform te ontwikkelen.

De voornaamste beperkingen van de techniek waren verbonden aan het gebruik van carboxypeptidasen. De enzymatische activiteit van carboxypeptidasen is immers sterk afhankelijk van zowel het af te splitsen C-terminaal aminozuur als van de volgorde van de aminozuren van het peptide, waardoor de snelheid waarmee de C-terminale degradatie gebeurde zeer sterk afhing van de aard van het peptide. Sommige aminozuren worden zelfs helemaal niet afgesplitst. We hebben nu een strategie voor chemische selectie ontwikkeld die het gebruik van carboxypeptidasen overbodig maakt. Door de homoserine lacton ringen gedeeltelijk te openen in een licht basische bufferoplossing worden alle interne en N-terminale peptiden tijdens MALDI TOF MS analyse geobserveerd als een doublet, met 18 Da massa verschil, terwijl het C-terminaal peptide als enige als singlet geobserveerd wordt. De sequentie van dit C-terminale fragment wordt vervolgens eenvoudig en met verhoogde gevoeligheid bepaald met behulp van MS/MS analyse. De techniek werd zowel voor eiwitten in oplossing als gel gescheiden eiwitten geoptimaliseerd. Om de throughput van de technologie te verhogen werd de staalvoorbereiding geïmplementeerd op een Tecan Freedom Evo 150 robot platform. Als proof-of-concept werden 96 2D-PAGE gescheiden eiwitten van de bacterie *Shewanella oneidensis*, zowel met de manuele CPase gebaseerde techniek, als de geautomatiseerde chemische selectie techniek geanalyseerd. Met de geautomatiseerde chemische selectie technologie werden drie keer meer eiwitten geïdentificeerd dan met de manuele CPase techniek, terwijl de staalvoorbereiding 10 maal sneller verliep.

De resultaten van de manuele chemische selectie techniek zoals toegepast op eiwitten in oplossing en gel gescheiden eiwitten werden in 2010 gepubliceerd in het tijdschrift Journal of Proteomics [16]. De automatisatie van de techniek en de resultaten van de vergelijkende studie op 96 *S. oneidensis* stalen werden in 2014 gepubliceerd in het open access tijdschrift EuPA Open Proteomics [17].

Bij deze chemische selectie technologie wordt een MALDI TOF/TOF MS instrument gebruikt voor de massaspectrometrische analyse. Uit analyse van de *S. oneidensis* dataset is gebleken dat slechts een beperkt aantal C-terminale peptiden een massa hebben die optimaal is voor een goede MS en MS/MS detectie. C-terminale CNBr peptiden zijn typisch te groot. Door aan het CNBr-reactiemengsel een kleine hoeveelheid KI toe te voegen treedt een splitsing op na zowel Met als Trp [18]. Tijdens deze oxidatie wordt Trp omgezet tot een Cγ-O-spirolactonderivaat. Omdat dit spirolactonderivaat chemisch gelijkaardig is aan het homoserinelactonderivaat, kon verondersteld worden dat de lacton functies zich in de chemische selectie technologie gelijk zullen gedragen. De reactieparameters voor de splitsingsreactie werden geoptimaliseerd en de reactieproducten en zijreacties gekarakteriseerd. Tijdens de splitsingsreactie worden eveneens alle disulfidebruggen verbroken en de cysteïnes geoxideerd tot cysteïnezuur. Hierdoor wordt het mogelijk om de gebruikelijke reductie-, alkylatie en ontzoutingsstap over te slaan. De gecombineerde partiële spirolacton- en homoserinelactonring opening werd getest bij twee eiwitten, waarbij het C-terminale peptide in het PMF mengsel kon geïdentificeerd worden en bij één eiwit de sequentie via *de novo* sequencing bepaald kon worden. Ondanks het feit dat tryptofaan ook een laag abundant aminozuur is (1.25 %), stijgt de theoretische aantal eiwitten waarvan de C-terminale sequentie kan worden bepaald met *de novo* sequentieanalyse met een kleine 10 % tot ongeveer 42 % van het totale aantal eiwitten in een bacterieel genoom.

De chemische selectie strategie kan enkel toegepast worden op eiwitten die vooraf met 2D-PAGE gescheiden zijn. 2D-PAGE is een methode met veel intrinsieke beperkingen, waardoor men steeds meer overgaat tot LC-MS gebaseerde terminale sequentie analyse technieken [12, 13, 19]. In dit proefschrift wordt een nieuwe LC-MS compatibele techniek voorgesteld. In onze methode worden de vrije carboxylgroepen van een eiwit (zijketens en C-terminus) gederivatiseerd met piperazine groepen, waarna de eiwitten gedigereerd worden met trypsine (of een protease met een andere specificiteit). Vervolgens worden de interne peptiden met een vrije carboxylgroup gekoppeld aan een resin, waarbij enkel het C-terminale peptide in oplossing blijft en geanalyseerd kan worden met (LC-)MS. Door de aanwezigheid van een piperazine groep op de C-terminus kunnen echte termini tijdens de MS analyse van vals positieven onderscheiden worden. De initiële resultaten, optimalisatie van derivatisatie en koppeling van interne peptiden aan de resin, worden hier voorgesteld.

Door de ontwikkeling van de chemische selectie techniek en de uitbreiding ervan met splitsing C-terminaal na methione en tryptofaan, is nu een platform ontwikkeld dat intrinsiek een eenvoudige, robuste en snelle C-terminale sequentie bepaling toelaat. Bovendien werd de basis gelegd voor een methode die niet langer afhankelijk is van 2D-PAGE technologie. Deze methoden kunnen in de toekomst gebruikt worden ter ondersteuning van genoomannotatieprojecten. De technieken kunnen ook leiden tot de identificatie van proteolytische processing van eiwitten, een proces waarvan eerder is aangetoond dat het een belangrijke impact kan hebben op de functie van een eiwit en in verband is gebracht met neurologische en hart-en vaatziekten [20].

References

[1] Mann, M. and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. *Nature biotechnology*, **21**, 255–261.

[2] Gupta, N., et al. (2007) Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome research*, **17**, 1362–1377.

[3] Castellana, N. and Bafna, V. (2010) Proteogenomics to discover the full coding content of genomes: A computational perspective. *Journal of proteomics*, **73**, 2124–2135.

[4] Venter, E., Smith, R. D., and Payne, S. H. (2011) Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *Plos One*, **6**.

[5] Maillet, I., Berndt, P., Malo, C., Rodriguez, S., Brunisholz, R. A., Pragai, Z., Arnold, S., Langen, H., and Wyss, M. (2007) From the genome sequence to the proteome and back: Evaluation of *E-coli* genome annotation with a 2-D gel-based proteomics approach. *Proteomics*, **7**, 1097–1106.

[6] Savidor, A., Donahoo, R. S., Hurtado-Gonzales, O., VerBerkmoes, N. C., Shah, M. B., Lamour, K. H., and McDonald, W. H. (2006) Expressed peptide tags: An additional layer of data for genome annotation. *Journal of proteome research*, **5**, 3048–3058.

[7] Samyn, B., Hardeman, K., Van der Eycken, J., and Van Beeumen, J. (2000) Applicability of the alkylation chemistry for chemical C-terminal protein sequence analysis. *Analytical chemistry*, **72**, 1389–1399.

[8] Zhong, H., Zhang, Y., Wen, Z., and Li, L. (2004) Protein sequencing by mass analysis of polypeptide ladders after controlled protein hydrolysis. *Nature biotechnology*, **22**, 1291–1296.

[9] Suckau, D. and Resemann, A. (2003) T(3)-sequencing: Targeted characterization of the N- and C-termini of undigested proteins by mass spectrometry. *Analytical chemistry*, **75**, 5817–5824.

[10] Sechi, S. and Chait, B. (2000) A method to define the carboxyl terminal of proteins. *Analytical chemistry*, **72**, 3374–3378.

[11] Kosaka, T., Takazawa, T., and Nakamura, T. (2000) Identification and C-terminal characterization of proteins from two-dimensional polyacrylamide gels by a combination of isotopic labeling and nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry. *Analytical chemistry*, **72**, 1179–1185.

[12] Dormeyer, W., Mohammed, S., van Breukelen, B., Krijgsveld, J., and Heck, A. J. R. (2007) Targeted analysis of protein termini. *Journal of proteome research*, **6**, 4634–4645.

[13] Van Damme, P., Staes, A., Bronsoms, S., Helsens, K., Colaert, N., Timmerman, E., Aviles, F. X., Vandekerckhove, J., and Gevaert, K. (2010) Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nature methods*, **7**, 512–515.

[14] Samyn, B., Sergeant, K., Castanheira, P., Faro, C., and Van Beeumen, J. (2005) A new method for C-terminal sequence analysis in the proteomic era. *Nature methods*, **2**, 193–200.

[15] Samyn, B., Sergeant, K., and Beeumen, J. V. (2006) A method for C-terminal sequence analysis in the proteomic era (proteins cleaved with cyanogen bromide). *Nature protocols*, **1**, 317–322.

[16] Moerman, P., Sergeant, K., Debyser, G., Devreese, B., and Samyn, B. (2010) A new chemical approach to differentiate carboxy terminal peptide fragments in cyanogen bromide digests of proteins. *Journal of proteomics*, **73**, 1454–1460.

[17] Moerman, P., Sergeant, K., Debyser, G., Timperman, I., Devreese, B., and Samyn, B. (2014) Automation of C-terminal sequence analysis of 2D-PAGE separated proteins. *EuPA open proteomics*, **3**, 250–261.

[18] Huang, S. and Huang, J. (1994) Cleavage of both tryptophanyl and methionyl peptide-bonds in proteins. *Journal of protein chemistry*, **13**, 450–451.

[19] Schilling, O., Barre, O., Huesgen, P. F., and Overall, C. M. (2010) Proteome-wide analysis of protein carboxy termini: C terminomics. *Nature methods*, **7**, 508–U33.

[20] Petrera, A., Lai, Z. W., and Schilling, O. (2014) Carboxyterminal protein processing in health and disease: key actors and emerging technologies. *Journal of proteome research*.

Summary

The achievements in genome research, e.g. the completion of the Human Genome Project, provided a treasure of valuable biochemical and biomedical information. In the meantime the focus shifted towards the biological interpretation of this genome information. Proteomics entails the large scale identification, quantification and characterization of the final gene products, i.e. the proteins and their post-translational modifications. Monitoring protein expression profiles and protein modifications remains a very challenging task because of the wide dynamic range of expressed levels of proteins and the variability of gene products due to the presence of splicing variants, N- and C-terminal processing, and co- and post-translational modifications (PTMs) [1]. Truncations of the nascent polypeptide chain at the N- or C-terminus are by far the most common types of PTMs found in proteins. Although these modifications are known to alter the biological activity of a protein, relatively little attention has been paid to the development of approaches for the systematic analysis of proteolytic processing events.

Systematic analysis of protein termini could also contribute to more accurate and reliable genome annotation. Currently, multiple algorithms for gene prediction and gene homology are combined to provide more reliable gene annotations. In proteogenomics, proteomics data is used as an additional factor to confirm or correct theoretically predicted genes [2–6]. Determining the exact C-terminus of the final protein product is one of the major challenges in this field.

Over the years, multiple methods have been developed to determine the C-terminal sequence of proteins. Initially chemical sequencing methods, complementary to Edman degradation, were developed. The derivatization of the much less chemically reactive carboxylic acid group proved to be exceedingly more challenging, limiting the sensitivity and robustness of these methods [7]. Later, several mass spectrometry based techniques were reported including limited protein hydrolysis [8] or the analysis of in-source decay fragments using a MALDI TOF mass analyzer equipped with a LIFT module [9]. In some techniques, C-terminal peptides were selectively enriched or labelled; e.g. enrichment using anhydrotrypsin columns [10] and labelling with ^{18}O isotopes during proteolysis [11]. Recently, a number of LC-MS based methods have been developed to determine protein termini. Ion exchange separations were used to enrich N-terminally blocked and C-terminal peptides [12], the COFRADIC technology was applied to specifically select N- and C-terminal peptide through diagonal chromatography [13]. None of

these methods entered into a mature state of wide scale usage, so there is still a demand for simple, sensitive and robust sequencing technologies.

The Laboratory for Protein Biochemistry and Biomolecular Engineering (L-ProBE) has a history in developing new methods for C-terminal sequence analysis. In 2005, a new method was reported in which carboxypeptidases (CPase) are used to selectively generate C-terminal sequence ladders in a mixture of cyanogen bromide (CNBr) cleaved proteins [14, 15]. During CNBr cleavage, all Met-Xxx peptide bonds are cleaved and all methionine residues are converted to homoserine lactone. During digestion with carboxypeptidase, only the original C-terminal fragment is accessible to enzymatic degradation and forms a ladder. The main goals of this project were to further improve the technique by eliminating some of its limitations and to automate the improved method, so that it can be applied to complex biological samples in a high-throughput setup.

We have now developed a strategy to chemically select the C-terminal peptide, eliminating the use of CPases. In this chemical selection method, all homoserine lactone residues, present in internal and N-terminal peptides generated by CNBr cleavage, are partially opened in a slightly basic buffer and are observed as a doublet, with 18 Da mass difference, during MALDI TOF MS analysis. This allows discriminating the C-terminal peptide, observed as singlet, which can then be identified by MS/MS. The protocol was optimized both for proteins in solution and gel-separated proteins. The sample preparation throughput was improved by implementing the technology on a Tecan Freedom Evo 150 robotic platform. As a proof-of-concept 96 2D-PAGE separated *Shewanella oneidensis* proteins were analyzed using both the manual CPase based protocol and the automated chemical selection procedure. We identified three times more proteins using the chemical selection technology, while sample preparation was 10 times faster.

The results obtained by applying the manual chemical selection technology on proteins in solution and gel-separated proteins were published in 2010 in Journal of Proteomics [16]. The implementation of the technology on an automated platform and the results of the differential study on 96 *S. oneidensis* samples were published in 2014 in the open access journal EuPA Open Proteomics [17].

In the chemical selection technology, a MALDI TOF/TOF instrument is used for MS analysis. The *S. oneidensis* dataset showed that only a limited number of C-terminal peptides have a mass in the optimal MS and tandem MS range. The peptides generated after CNBr cleavage are often relatively large. Adding KI to the CNBr cleavage mixture induces cleavage C-terminal of both methionine and tryptophan [18]. During this oxidative halogenation Trp is converted to a Cγ-O-spirolactone tryptophan. The structural resemblance between the reaction products

offered the likelihood of a similar behavior during chemical selection incubations. We optimized the protocol and characterized the reaction products and the side reactions. During protein cleavage, also disulfide bridges were cleaved and cysteine residues were oxidized to cysteic acid. This allows to eliminate the standard reduction, alkylation and desalting step. Using two test proteins, we were able to identify both C-terminal peptides in a PMF mixture after combined partial spiro- and homoserinelactone ring opening. In one of the cases we were also able to obtain the C-terminal sequence after *de novo* interpretation of the fragment spectrum. Even though tryptophan is also a low abundant amino acid (1.25 %), the theoretical number of proteins that can be identified by *de novo* sequencing is raised by 10 % to around 42 % of the total protein complement of a bacterial genome.

The chemical selection technology requires proteins to be separated by 2D-PAGE prior to analysis. To circumvent the limitations associated with 2D-PAGE, recently, several LC-MS-based terminal sequencing technologies have been presented [12, 13, 19]. We also present a new LC-MS compatible technology. In our method free carboxyl groups of a protein (both side chains and C-terminus) are derivatized using piperazine groups followed by digestion with trypsin (or any other protease with different specificity). Next, the internal peptides, containing a free carboxylgroup are coupled to a resin. The C-terminal peptide remains in solution and can be analyzed using (LC-)MS. The presence of the piperazine group at the C-terminus of the proteins allows differentiating real protein termini from false positives. The initial results, optimizing the derivatization reaction and coupling of internal peptides to the resin, are presented here.

By combining the initial chemical selection technology with the cleavage C-terminal of methionine and tryptophan, a platform has been developed that allows C-terminal sequence analysis in a simple, robust and high-throughput way. In the future this sequencing platform can be applied to support genome annotation projects. It can also be used to study proteolytical processing events, a process that can significantly alter the function of a protein and has recently been connected to neurological or cardiovascular diseases (Petrera 2014). To circumvent the limitations associated with 2D-PAGE separations the initial results of an alternative LC-MS compatible sequencing strategy has been presented [20].

References

[1] Mann, M. and Jensen, O. N. (2003) Proteomic analysis of post-translational modifications. *Nature biotechnology*, **21**, 255–261.

[2] Gupta, N., et al. (2007) Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome research*, **17**, 1362–1377.

[3] Castellana, N. and Bafna, V. (2010) Proteogenomics to discover the full coding content of genomes: A computational perspective. *Journal of proteomics*, **73**, 2124–2135.

[4] Venter, E., Smith, R. D., and Payne, S. H. (2011) Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *Plos One*, **6**.

[5] Maillet, I., Berndt, P., Malo, C., Rodriguez, S., Brunisholz, R. A., Pragai, Z., Arnold, S., Langen, H., and Wyss, M. (2007) From the genome sequence to the proteome and back: Evaluation of *E-coli* genome annotation with a 2-D gel-based proteomics approach. *Proteomics*, **7**, 1097–1106.

[6] Savidor, A., Donahoo, R. S., Hurtado-Gonzales, O., VerBerkmoes, N. C., Shah, M. B., Lamour, K. H., and McDonald, W. H. (2006) Expressed peptide tags: An additional layer of data for genome annotation. *Journal of proteome research*, **5**, 3048–3058.

[7] Samyn, B., Hardeman, K., Van der Eycken, J., and Van Beeumen, J. (2000) Applicability of the alkylation chemistry for chemical C-terminal protein sequence analysis. *Analytical chemistry*, **72**, 1389–1399.

[8] Zhong, H., Zhang, Y., Wen, Z., and Li, L. (2004) Protein sequencing by mass analysis of polypeptide ladders after controlled protein hydrolysis. *Nature biotechnology*, **22**, 1291–1296.

[9] Suckau, D. and Resemann, A. (2003) T(3)-sequencing: Targeted characterization of the N- and C-termini of undigested proteins by mass spectrometry. *Analytical chemistry*, **75**, 5817–5824.

[10] Sechi, S. and Chait, B. (2000) A method to define the carboxyl terminal of proteins. *Analytical chemistry*, **72**, 3374–3378.

[11] Kosaka, T., Takazawa, T., and Nakamura, T. (2000) Identification and C-terminal characterization of proteins from two-dimensional polyacrylamide gels by a combination of isotopic labeling and nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry. *Analytical chemistry*, **72**, 1179–1185.

[12] Dormeyer, W., Mohammed, S., van Breukelen, B., Krijgsveld, J., and Heck, A. J. R. (2007) Targeted analysis of protein termini. *Journal of proteome research*, **6**, 4634–4645.

[13] Van Damme, P., Staes, A., Bronsoms, S., Helsens, K., Colaert, N., Timmerman, E., Aviles, F. X., Vandekerckhove, J., and Gevaert, K. (2010) Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nature methods*, **7**, 512–515.

[14] Samyn, B., Sergeant, K., Castanheira, P., Faro, C., and Van Beeumen, J. (2005) A new method for C-terminal sequence analysis in the proteomic era. *Nature methods*, **2**, 193–200.

[15] Samyn, B., Sergeant, K., and Beeumen, J. V. (2006) A method for C-terminal sequence analysis in the proteomic era (proteins cleaved with cyanogen bromide). *Nature protocols*, **1**, 317–322.

[16] Moerman, P., Sergeant, K., Debyser, G., Devreese, B., and Samyn, B. (2010) A new chemical approach to differentiate carboxy terminal peptide fragments in cyanogen bromide digests of proteins. *Journal of proteomics*, **73**, 1454–1460.

[17] Moerman, P., Sergeant, K., Debyser, G., Timperman, I., Devreese, B., and Samyn, B. (2014) Automation of C-terminal sequence analysis of 2D-PAGE separated proteins. *EuPA open proteomics*, **3**, 250–261.

[18] Huang, S. and Huang, J. (1994) Cleavage of both tryptophanyl and methionyl peptide-bonds in proteins. *Journal of protein chemistry*, **13**, 450–451.

[19] Schilling, O., Barre, O., Huesgen, P. F., and Overall, C. M. (2010) Proteome-wide analysis of protein carboxy termini: C terminomics. *Nature methods*, **7**, 508–U33.

[20] Petrera, A., Lai, Z. W., and Schilling, O. (2014) Carboxyterminal protein processing in health and disease: key actors and emerging technologies. *Journal of proteome research*.

Part I

General Introduction

Chapter 1

Proteomics and mass spectrometry

1.1 Introduction

The Human Genome Project was completed a decade ago [1] and resulted in 20.135 annotated unique genes, a lot less than the anticipated 100.000 [2]. This led to a paradox; why are there so few genes in the human genome compared to organisms that are considered less complex? The cabbage genome, for instance, encodes twice as many genes compared to humans, 41.174. Many other organisms have a similar number of genes as humans. The *Drosophila* genome contains 15.016 genes and *Caenorhabditis elegans* contains 20.470, so where does the extra information come from that makes humans distinct from other organisms [3]?

It soon became clear that organism complexity is generated by a complex proteome rather than by a complex genome. The estimated number of protein forms encoded by these genes is two to three orders of magnitude higher. The proteome is defined here as the time- and cell-specific protein complement of the genome. It encompasses all proteins that are expressed in a cell at a certain moment, including isoforms and protein modifications. Whereas the genome is constant for one cell, largely identical for all cells of an organism, and does not vary a lot within species, the proteome is very dynamic with time and in response to external factors, and differs substantially between cell types and subcellular localizations [4]. At the DNA, RNA and protein levels, complexity can arise from allelic variations, from alternative splicing of RNA transcripts and from many post-translational modifications, respectively. These events create distinct protein molecules that modulate a wide variety of biological processes, from cell signaling inside or between cells to gene regulation and activation of protein complexes [5]. The successor of the Human Genome Project, the Human Proteome Project was launched in 2010 by HUPO (Human Proteome Organization) to identify and understand the function of all proteins in the human body and to map the human proteome on a disease- and chromosome-centric basis [3].

1.2 Information flow in Biology

The central dogma of molecular biology describes, in its simplest form, the information flow in biological systems from deoxyribonucleic acid (DNA) to DNA during replication or DNA via ribonucleic acid (RNA) to a synthesized protein during respectively transcription and translation [6] (Figure 1.1). While DNA forms a catalogue of the biomolecular information of an organism, mRNA serves as a messenger to transfer that information in the cell. Proteins form the functional entities of an organism, although several processes are known where RNA is the effector molecule, e.g. tRNA, rRNA and different forms of small RNAs. Both transcription and translation are regulated quantitatively and qualitatively, and together provide a quick protein response to environmental changes. A single DNA sequence can encode multiple proteins due to a variation in translation start and stop sites, translational frame shifting, and in eukaryotes also by alternative splicing, the use of alternative promoters and RNA-editing. Mainly in eukaryotes, the proteins can also undergo a wide range of co- and post-translational modifications. The different processing steps and quantitative regulation steps explain why no simple correlation can be made between the expression levels of genes and proteins [7, 8].

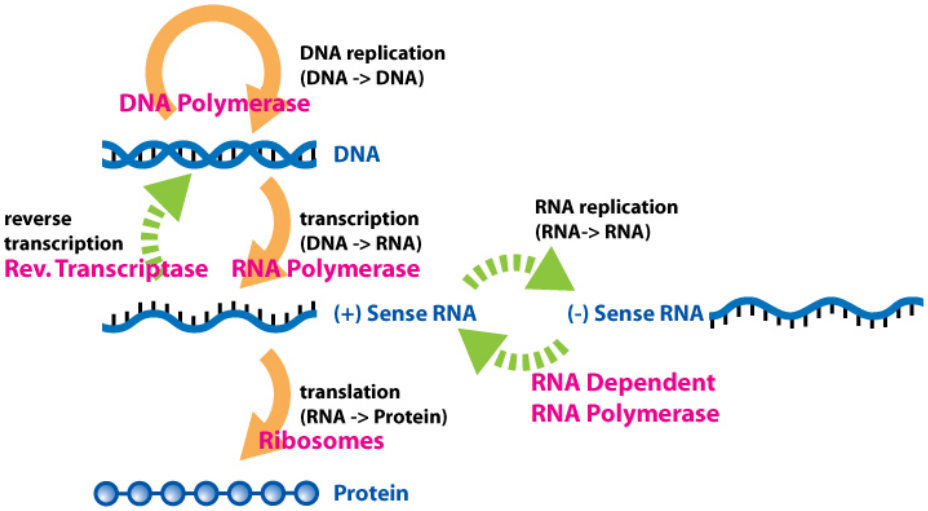


Figure 1.1: Overview of the different information flows according to the central dogma of molecular biology. Flows in green only appear in certain organisms. Reverse transcription is typical for retroviruses, like HIV. RNA replication is common in RNA viruses [9].

Several different scientific disciplines focus on the different steps in the central dogma, genomics (study of the genome) [10, 11], transcriptomics (study of transcriptome) [12, 13], proteomics (study of the proteome) and metabolomics (global study of the metabolites) [14, 15]. By combining data from different omics technologies, system biologists are able to get a holistic view on the complexity of biological systems.

1.3 Proteomics: definition, challenges and technology

The original definition of the term "protein complement encoded by a genome" does not reflect the dynamic properties of what is currently known as the proteome [16]. A more comprehensive definition could be "proteomics is the study of all proteins expressed at a given moment under given circumstances by a specific type of cell or tissue".

The enormous complexity of the proteome makes the development of appropriate technology for its analysis extremely challenging. Proteins are not only distributed over all cell structures, some are water-soluble cytoplasmic proteins and others are membrane-associated or integral membrane proteins, requiring different extraction and purification methods. Protein expression levels differing by 12 orders of magnitude in biological samples have been observed, asking for techniques with a very large dynamic range [17]. Post-translational modifications can result in significant activity difference, but often affects only a small fraction of the total amount of an expressed protein. Unlike for oligonucleotides, no amplification technique is available for proteins to overcome the problem of detection limits. Proteomic methods are typically biased to the more abundant proteins. Several affinity based depletion techniques have been reported. Up to 20 of the most abundant proteins from blood serum samples can be depleted, but even if the depletion is 99.9% successful, those 20 proteins are still 10^{9th}-fold more abundant than many of the important signaling proteins [18]. Therefore, extensive separations prior to analysis are necessary to characterize the total protein complement of a biological sample.

The introduction of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) in 1975 by O'Farrell can be considered as the birth of proteomics [19]. Proteins are first separated according to their pI value during an iso-electric focusing step in the first dimension. In the orthogonal second dimension the proteins are separated on the basis of their molecular weight by sodium dodecylsulfate polyacrylamide gel electrophoresis. Although the power of large 2D-PAGE gels has been demonstrated by separating thousands of proteins, the approach has some serious limitations [20, 21]. Hydrophobic and low abundant proteins, as well as those with extreme pI values are often missed. In addition, the technique is labor intensive and is not amenable for automatic high-throughput analysis. Alternatively, complex protein mixtures can be separated using liquid chromatography and capillary electrophoresis. Although multiple types of resins are available for protein chromatography, none provides sufficient resolution to be considered as an alternative to 2D-PAGE for protein-based separations [22].

Two fundamental breakthroughs facilitated protein identification of unknown protein mixtures, i.e. the complete sequencing of genomes and the development of two mass spectrometric (MS) ionization techniques. Prior to these achievements separated proteins were identified using Edman sequencing, a method that lacks the sensitivity and throughput to systematically

characterize 2D-PAGE separated protein spots.

Due to these fundamental developments, proteomic approaches were focusing on mass spectrometry based gel-free peptide-centered methods. Techniques in which peptides are separated and analyzed using MS are called 'bottom-up' or 'shotgun' proteomics approaches, as opposed to 'top-down' approaches where intact proteins are separated and characterized by MS. In shotgun proteomics experiments, proteins are enzymatically cleaved and the resulting complex peptide mixture is chromatographically separated prior to MS analysis. These separation steps are necessary to reduce the sample complexity, because even high-resolution mass spectrometers are unable to handle the large amount of peptides generated after digestion of thousands of proteins. Trypsin is typically used as protease in bottom-up approaches. It cleaves C-terminal of Arg and Lys residues, and due to the natural abundance of these two amino acids in proteins, peptides are generated that fit the m/z range of mass analyzers [23]. Recently the use of multiple proteases has gained attention (LysC, ArgC, AspN, GluC). They are complementary to trypsin as they increase the proteome coverage of the identified peptides [24]. In order to separate these complex mixtures a combination of multiple and/or different LC methods is required. The group of John Yates III first developed an online method coupling two-dimensional liquid chromatography to tandem mass spectrometry. In this method a microcapillary column was packed with two independent chromatographic phases, strong cation exchange phase (SCX) and C18 reverse phase (RP) (Figure 1.2). The technique was called Multidimensional Protein Identification Technology (MudPIT)[25]. Together with the protein identification algorithm SEQUEST, it constituted the first automated platform for proteome analysis [26, 27]. Aside from SCX-RP, several other 2D-LC setups have been found to provide sufficient orthogonality [28]. Combining two reverse phase separations, the first one at pH 10, the second one at pH 3, have been described [29]. Although peptide hydrophobicity remains the major separation parameter, sufficient orthogonality is achieved using RP-RP LC. Depending on the conditions (pH <pI or pH >pI) certain residues (Arg, His, Lys, Asp, Glu) either gain a charge and become hydrophilic, or lose their charge and become hydrophobic. During 'off-line' coupled separations, the eluent of the first chromatographic separation is collected and re-injected for the second separation. In 'online' techniques, the samples are directly transferred from the first dimension column to the second dimension column.

1.4 Mass spectrometry

Mass spectrometry has become the method of choice for the analysis of complex protein samples. MS-based proteomics is a discipline made possible by the availability of gene and genome sequence databases and technical and conceptual advances in many areas, most notably the

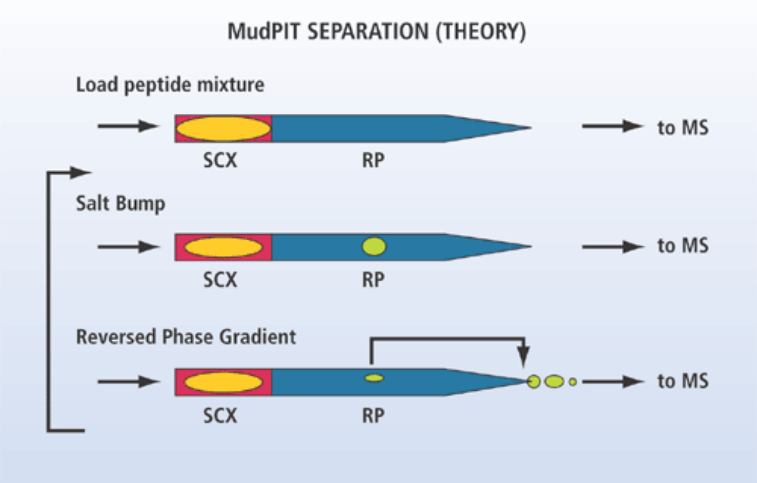


Figure 1.2: A schematic overview of the MudPIT approach. Both SCX and RP resin are loaded into an ESI needle. By alternating increasingly stronger salt pulses with reverse phase gradients, peptides are eluted from the material [30].

discovery and development of soft ionization techniques, as recognized by the 2002 Nobel Prize in chemistry [31].

A mass spectrometer generally consists of three compartments, an ion source, a mass analyzer and a detector. In the ion source, analyte molecules are charged and transferred into the gas-phase before being passed onto the mass analyzer. Using an electrical or magnetic field, a mass analyzer separates the gas-phase ions according to their mass to charge (m/z) ratio. A detector registers the ions in a quantitative way.

Multiple mass analyzers can be put in series, usually with a fragmentation unit in between them. This setup allows to perform so called 'tandem MS'. The precursor ion is selected by the first mass analyzer, is fragmented in the fragmentation unit, and the second mass analyzer separates the newly formed fragment ions. A peptide sequence can be determined from the generated MS/MS spectrum. In hybrid instruments multiple types of mass analyzers are combined, e.g. Q-TOF.

Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) are the two techniques most commonly used to volatize and ionize the proteins or peptides for mass spectrometric analysis [32–34]. ESI ionizes the analytes out of a solution and is therefore readily coupled to liquid-based (for example, chromatographic and capillary electrophoresis) separation tools. MALDI sublimates and ionizes the samples out of a dry crystalline matrix using laser pulses. MALDI-MS is normally used to analyze relatively simple peptide mixtures, whereas

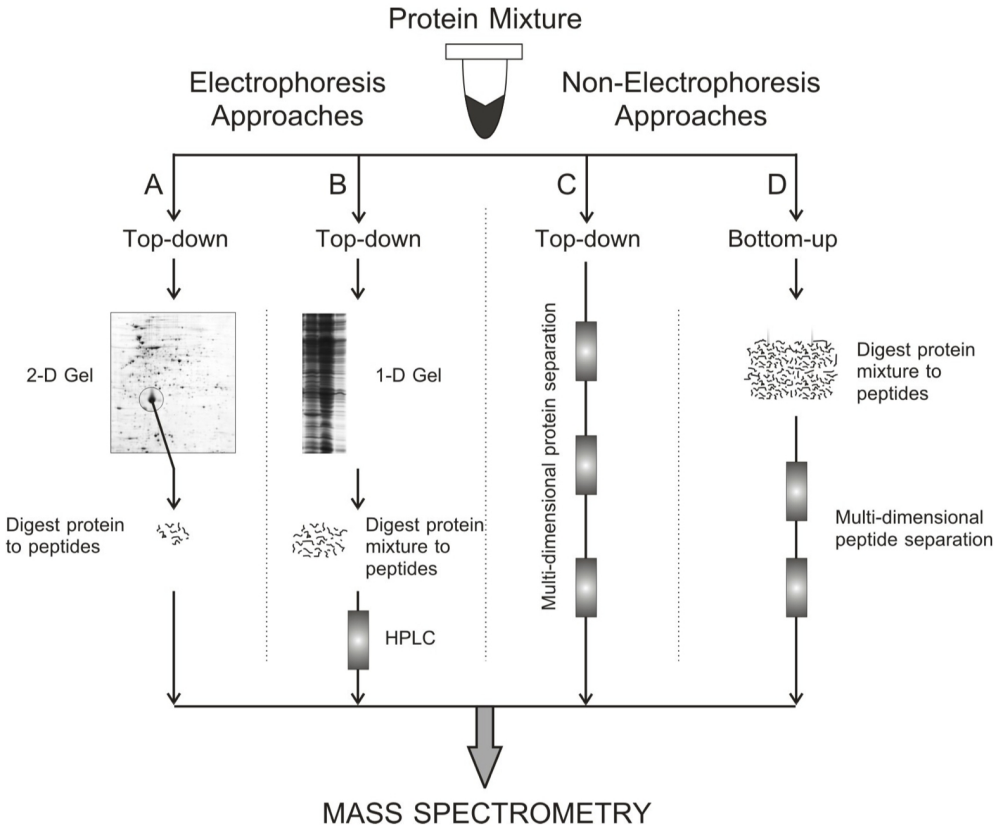


Figure 1.3: Workflow of a typical proteomic experiment. First the proteins are extracted from a sample and subjected to fractionation, before being enzymatically digested into a peptide mixture and identified by mass spectrometry. This reduction in complexity can either be achieved by using gel-based separation methods, such as two-dimensional gel electrophoresis (A) and geLC (B), or by using multiple HPLC separations of the proteins (C) or peptides (D) before MS identification.

integrated liquid-chromatography ESI-MS systems (LC-MS) are preferred for the analysis of complex samples. There are five basic types of mass analyzers currently used in proteomics research: the ion-trap, time-of-flight (TOF), quadrupole, orbitrap and Fourier transform ion cyclotron resonance (FT ICR-MS) analyzers. They are very different in design and performance, each with its own strengths and weaknesses [35].

AB Sciex 4800 plus MALDI TOF/TOF

All our experimental work was designed for and performed on a MALDI TOF/TOF instrument. Therefore, in this introduction we restrict the technical part on an outline of this specific set-up. In our work, an Applied Biosystems SCIEX 4800 Plus Proteomic analyzer was used (Figure

1.4). The MS was operated in two different modes, the reflector MS and the tandem MS mode. In the tandem MS mode, the MALDI-generated analyte ions are separated in the first TOF analyzer. A time-ion selector is used to isolate and transfer the precursor ion into the collision cell, where the selected ion is fragmented. The fragment ions are then reaccellerated by the second source and separated by the second TOF analyzer. In the reflector MS mode the analyte ions are not fragmented and both TOF analyzers are used in series, resulting in a longer flight path and higher resolution. The different elements of a MALDI TOF/TOF instrument are discussed in more detail below.

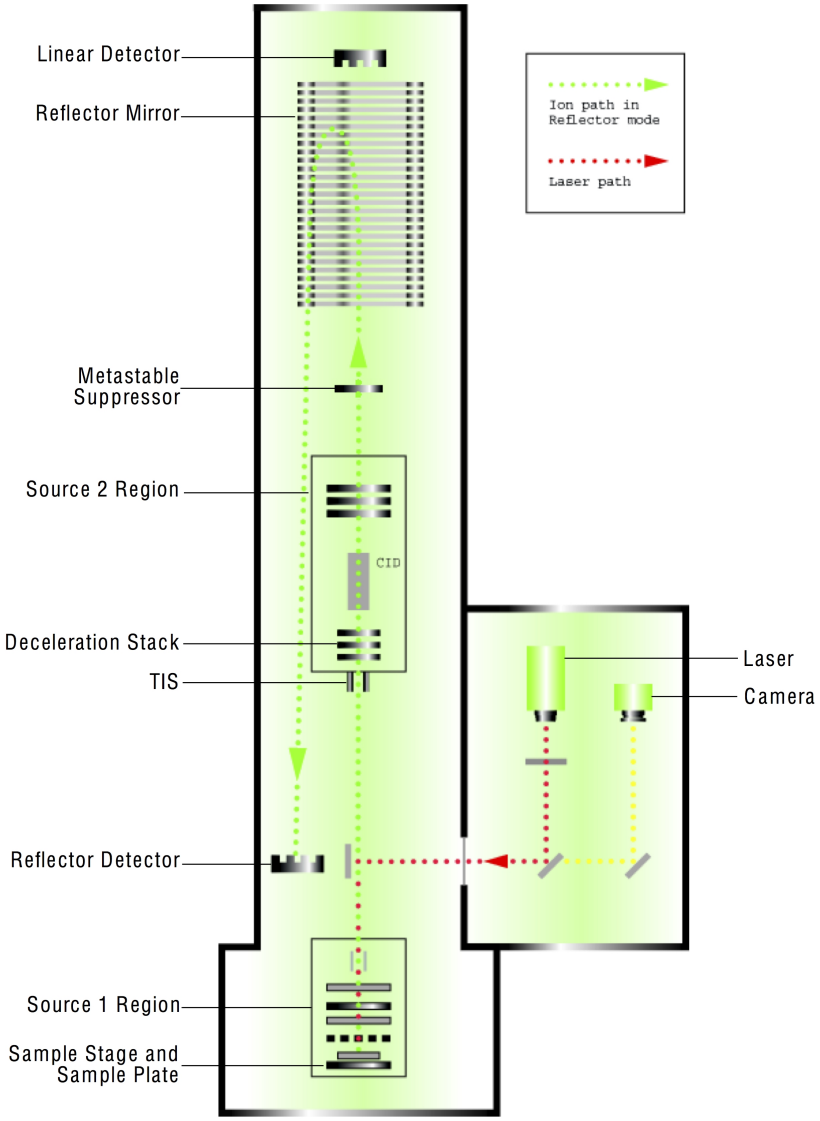


Figure 1.4: Schematic overview of the 4800 Plus MALDI TOF/TOF analyzer.

1.4.1 Matrix-Assisted Laser Desorption/Ionization (MALDI)

In the Matrix-Assisted Laser Desorption/Ionization (MALDI) process, the sample is mixed with a solution of low-molecular weight matrix molecules and spotted on a solid metallic MALDI target plate. The matrix and analyte molecules co-crystalize on the plate [36]. Ionization is achieved by irradiating the target with a pulsed laser, typically an Nd-YAG or N_2 laser generating light in the UV-spectrum. The matrix molecules (e.g. α-cyano-4-hydroxycinnamic acid for peptide analysis and 3,5-dimethoxy-4-hydroxycinnamic acid for protein analysis [37, 38]) have an absorption maximum at or near the wavelength of the laser and are sublimated from the target plate along with the analyte molecules. Analyte molecules are ionized in the gas phase by proton transfer from the matrix to the analyte. Although well studied, the precise nature of the ionization process in MALDI remains under discussion [39–41]. After ionization the, mostly single, charged ions are accelerated by a voltage applied to a grid and extracted from the source towards the mass analyzer (Figure 1.5).

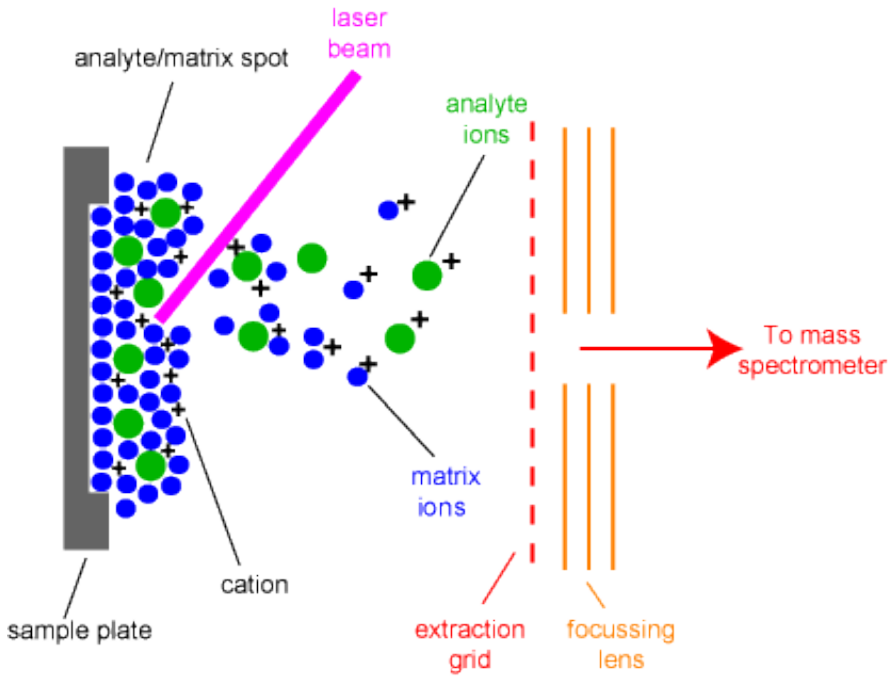


Figure 1.5: Principles of the Matrix-Assisted Laser Desorption/Ionization (MALDI) process.

1.4.2 Time of Flight (TOF) Analyzer

A Time of Flight analyzer is essentially a long vacuum tube with an ion-source at one end and a detector at the other (Figure 1.6) [42]. The ions, typically generated by the MALDI process, are accelerated towards the field-free region by an electric field V_e generated by the applied grid voltage at the source. The kinetic energy ($E_{kin} = \frac{1}{2}mv^2$) of an ion in the field-free TOF

tube equals the acceleration energy it has received during the extraction ($E_{acc} = zV_e$). The flight time of an ion (t) therefore depends on the mass-to-charge ratio (m/z) of the ion and the length of the flight tube (x), according to formula 1.1;

$$t = \left(\frac{m}{z} \cdot \frac{x^2}{2V_e} \right)^{\frac{1}{2}} \tag{1.1}$$

According to this equation, ions with a lower m/z ratio reach the detector earlier than ions with a higher m/z ratio, so separating them. Despite the simple concept, early TOF analyzers achieved poor resolution. As not all ions are formed simultaneously and at the same distance from the grid, they obtain a different kinetic energy, causing a broadening effect. This can be countered by adding a reflector to the flight tube and by so called 'delayed extraction' of the ions from the source [43–45]. The use of a reflector unfortunately lowers the sensitivity and imposes an upper m/z limit [46].

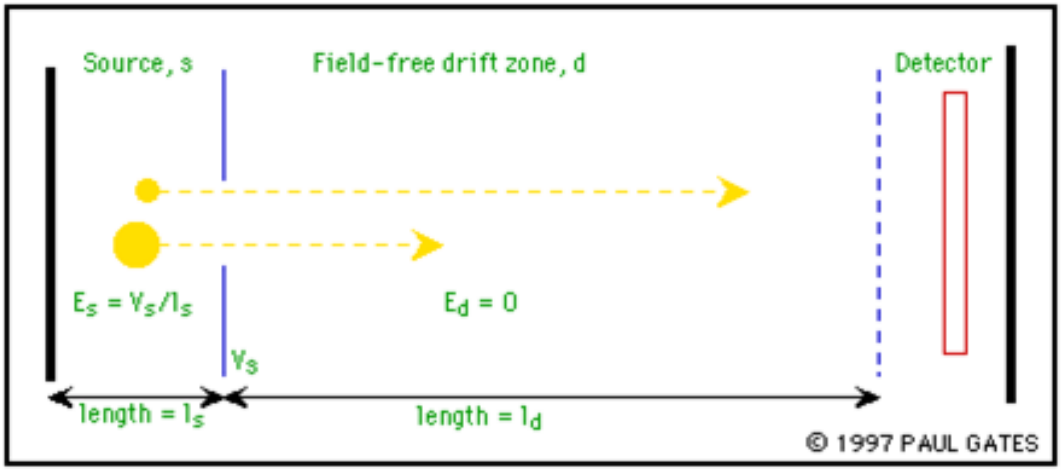


Figure 1.6: Overview of a TOF analyzer.

1.4.3 Collision cell

In a collision cell peptide ions are typically fragmented by collision-induced dissociation (CID), also referred to as collisionally activated dissociation (CAD). The precursor ions enter the collision cell at high kinetic energy and collide with neutral gas (He, N_2, Ar or Air). The loss of kinetic energy during the impact is converted into internal energy, resulting in bond cleavage and the formation of smaller fragment ions. On the AB MALDI TOF/TOF, depending on the mode, the collision cell is operated at 7 or 6 kV, respectively causing a 1 or 2 kV potential difference between the source (8 kV) and the collision cell. This potential difference causes the analyte molecules to fragment. Additionally neutral gas can be added to the collision cell, mainly resulting in the fragmentation of side chains and the formation of internal peptide ions.

Fragmentation mainly occurs along the peptide backbone, and the most labile bond in gas phase peptides is the peptide bond. The different fragment ions formed are named according to a standardized nomenclature (Figure 1.7)[47]. The generated fragment ions form a (sometimes interrupted) series of masses, each representing a peptide fragment differing from the next by one amino acid. As such, peptide sequences can be determined from the tandem mass or MS/MS spectra.

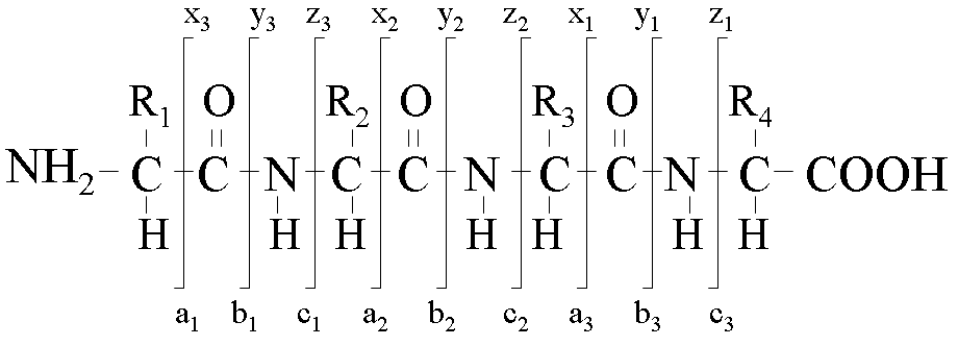


Figure 1.7: Nomenclature for the product ions generated after peptide fragmentation by tandem mass spectrometry. E.g. when the peptide bond is fragmented, the peptide fragement ion containing the N-terminus is called a b-ion, the fragment ion containing the C-terminus is called a y-ion.

1.5 Identification of proteins from peptide maps and MS/MS data

Besides *de novo* sequencing, where sequence tags are (manually) derived from tandem MS spectra and searched against protein databases using BLAST algorithms, two strategies are commonly used to identify proteins from MS-generated data, Peptide Mass Fingerprint (PMF) and Peptide Fragment Fingerprint (PFF). Both techniques require a protein database of the analyzed species or at least of a very closely related species. During PMF, a 2D-PAGE or chromatographically separated protein is proteolytically digested and the resulting peptide mixture is analyzed by MS. Every protein digested by an enzyme with known specificy will produce a specific series of peptide masses, called a 'mass fingerprint'. A protein can be identified by matching the experimentally obtained mass list against *in silico* generated PMFs of all proteins in the database. A scoring algorithm ranks the hits according to plausibility. Today, different database search algorithms are available that are differing mostly at the level of the scoring algorithm; examples are: MASCOT [48], SEQUEST [49], X!Tandem [50] and Paragon [51].

Similarly, in peptide fragment fingerprinting (PFF) the experimental MS/MS spectra of peptides are searched against a protein database. The search algorithm first selects all peptides in the database that are isobaric to the precursor after cleavage with a selected enzyme. These peptides are then *in silico* fragmented and the resulting peptide fragment fingerprints are compared with the experimental MS/MS spectrum. The most commonly used PFF search algorithms are MASCOT [48] and SEQUEST [49]. PFF analysis is used in shotgun proteomics experiments because PMF information is lost by the enzymatic cleavage of the proteins carried out prior to separation.

1.6 Alternative fragmentation techniques and Top-down Proteomics

Besides the very common combination of CID with an ion-trap or TOF analyzer, other tandem MS techniques are available for protein identification and characterization. Although these techniques can be used to characterize peptide mixtures, their main application lies in top-down proteomics studies. Currently both the FT-ICR and orbitrap analyzers are capable of determining the mass of intact proteins with high resolution and mass accuracy [52–54]. Although the intact protein mass can be used to confirm the structure of proteins with a known sequence, the information remains insufficient to identify unknown proteins and proteins containing different co- and post-translational modifications.

Fragmentation of the intact protein and interpretation of the fragmentation spectra provides additional information than the determination of the intact mass alone. To obtain this information, electron capture dissociation (ECD) and electron transfer dissociation (ETD) are two fragmentation techniques often implemented in high resolution mass analyzers. During ECD the analyte ions are bombarded with low energy electrons in the ICR cell. The interaction of the multiple charged analyte ions and the electrons results in the formation of odd electron ions that fragment into product ions. A similar process occurs in ETD, where radical anions of anthracene and azobenzene are used to transfer an electron to the multiple charged analyte ions. The ETD process is typically performed in a linear ion trap coupled to a high resolution analyzer. Whereas during CID, peptide bonds are fragmented after introduction of vibrational energy and intramolecular rearrangements, while ECD and ETD use a faster, non-ergodic process to fragment bonds more uniformly along the peptide backbone and predominantly generate c- and z-ions as opposed to b- and y-ions in CID [55–59].

In bottom-up proteomics approaches, C-terminal peptides are often not observed. Internal tryptic peptides always contain at least one positive charge, C-terminal peptides are often very small or have no positivly charged residue, limiting their detection in MS. The uniform frag-

mentation along the protein backbone by ECD and ETD allows deriving continuous sequence information, including the terminal regions, from tandem MS spectra. Some post-translational modifications, such as phosphorylations and glycosylations, are typically lost during the CID fragmentation process, but remain intact during ETD and ECD fragmentation. Multiple tandem mass analyzers have been used to obtain post-translational modification and terminal sequence information of small proteins; ESI-ECD-FTICR [55, 60, 61], ESI- Infrared multiphoton dissociation (IRMPD)-FTICR [62], ESI-IRMPD+ECD-FTICR [63–65], ESI-ETD-LTQ [66]. Besides the described post-source fragmentation techniques, also in-source fragmentation techniques have been applied to study protein sequences; for example the T3-sequencing approach on a MALDI LIFT-TOF/TOF [67, 68].

Most of the techniques can only be applied on relatively large amounts of small purified proteins and generate very complex datasets. Currently, the high cost and the difficulties accompanied with the purification of intact proteins prevents top-down approaches to compete with the existing bottom-up approaches. We will therefore focus on bottom-up approaches.

1.7 Quantification

Besides proteomic profiling, where a list of identified proteins is generated, most proteomic studies compare multiple environmental or disease states of an organism. The oldest differential proteomic method uses 2D-PAGE profiles to determine changes in protein abundance by comparing spot intensities. The intensities of Coomassie- or silver-stained protein spots can be compared, or proteins can be differentially stained with fluorescent dyes prior to 2D-PAGE separation (difference gel electrophoresis, DIGE) to determine the relative protein abundance in multiple samples using protein image analysis software [69].

Bottom-up MS-based proteomics, GeLC-MS/MS and MudPIT, can also be employed to determine differences in protein expression levels across samples. Relative or absolute quantification of peptides involves either label-free or label-incorporated approaches to discern differences in protein abundance among different biological conditions. Many reviews discuss the different label and label-free methods currently available; an overview is given in Table 1.1 [70–73].

Table 1.1: Overview of the most important and popular MS-based quantification methods in proteomics.

Method	Principle	Quantification	Advantages	Disadvantages	Ref.
ICAT	chemical labeling, cysteine specific	relative	enrichment of labeled peptides (reduced sample complexity)	no quantification of proteins containing no cysteine	[74]
iTRAQ	chemical labeling with isobaric tags, amine specific, quantification in MS/MS	relative	multiplex (up to eight samples), quantification of proteins in tissue	labeling efficiency needs to be checked, specific software for data analysis is required	[75]
^{15}N-labeling	metabolic labeling during cell growth	relative	complete introduction of stable isotopes of amino acids (labeling efficiency nearly 100%)	complex data analysis, extremely enriched nitrogen is needed	[76]
SILAC	metabolic labeling during cell growth	relative	complete introduction of stable isotopes (labeling efficiency nearly 100%)	no labeling in tissue	[77]
^{18}O-labeling	introduction of ^{18}O during enzymatic hydrolysis	relative	cheap reagents, simple labeling protocol	incomplete labeling complicates data analysis	[78–80]
AQUA	addition of stable isotope labeled standard peptides	absolute	absolute quantification of proteins/peptides in complex mixtures	for a small set of target proteins only	[81–83]
spectral count	quantification on the number of acquired MS/MS spectra	relative	no labeling required	semi-quantitative	[84]
ion counting	quantification on the ion intensity	relative	no labeling required	very high reproducibility of LC and MS is required	[85]

1.8 Terminomics

C- and N-terminomics can be defined as the proteome wide-analysis of protein C- and N-termini and their modifications.

In solution due to solvation, most protein termini are exposed at the protein surface and often serve as recognition sites for receptors. Several protein domains have been reported to specifically interact with protein C-terminal regions, e.g. PDZ and tetratricopeptide repeat (TPR) domains [86, 87]. The interactions of these protein domains with various terminal epitopes plays an important role in a broad range of physiological functions. Some of these functions include protein trafficking, subcellular anchoring of proteins, targeted protein degradation, and formation of macromolecular complexes. These interactions are often regulated by post-translational modifications (PTM) of the terminal regions [88].

While N-termini, e.g. due to the presence of an initiator methionine and the presence of a signal sequence for protein targeting, are often redundant, C-termini are very specific. Wilkins *et al.* evaluated the possibility to identify 2D separated proteins on the basis of short N- or C-terminal protein tags. C-terminal tags of 4 amino acids were found to be unique in up to 97% of the proteins, depending on the species studied [89]. While studying terminal sequences of all open reading frames in yeast, Li *et al.* showed that both previously characterized and new terminal sequences are conserved with the same frequency as functionally important, experimentally confirmed signals. This was also found in other organisms. It was shown that proteins can even be functionally classified based on their termini [88, 90].

Several modifications specific to protein termini are reported in Unimod, an online database of known protein modifications (www.unimod.org) [91]. These modifications are known to change protein interactions and alter the turn-over rate of proteins, e.g. causing deviation of the N-end rule that links protein half-life to the N-terminal amino acid [92]. 11 amino-terminal PTMs are defined including acetylation, mono-, di- and trimethylation, formylation, carbamylation, succinylation, cyclization, propionylation, palmitoylation and myristoylation. N-acetylation and cyclization have been studied in depth at the mechanistic and proteome-wide level [93–95]. The C-terminus of a protein is inherently less reactive than the N-terminus, resulting in fewer PTMs. C-methyl-esterification is the most frequently annotated modification of the carboxyl terminus, but with only 17 cases reported in human, 6 in mouse and 7 in yeast, it remains underexplored [96]. C-terminal isoprenylation, cholesterol-esterification and addition of glycophosphatidylinositol (GPI) anchors are involved in membrane targeting and trafficking. The limited number of described and observed C-terminal modifications might not reflect reality, but may rather be due to the lack of appropriate technologies [97].

The 'Termini-oriented protein Function Inferred Database' (TopFIND http://clipserve.clip.ubc.ca/topfind) acts as central repository and information resource to combine the information on protein termini, limited proteolysis and protein interactions via termini [98].

1.9 Proteogenomics

Proteogenomics has emerged as a field at the intersection of genomics and proteomics. The aim of the field is to improve and verify gene prediction and gene annotations using proteomic data. Several bioinformatics tools have been developed to improve the matching of MS/MS derived peptide sequences to genomic loci in pro- and eukaryotes [99–104]. Proteomic data has been used to confirm translation, determine reading-frames, identify gene and exon boundaries, provide evidence for post-translational processing, identify splice-forms including alternative splicing, and predict completely novel genes [99].

Traditionally, proteomic studies could only be performed on organisms with a sequenced genome, or on evolutionary close relatives. Nowadays, using next-generation sequencing, draft genomes can be generated at low cost for any organism. Proteogenomic software tools allow to identify tandem MS data using these draft genomes, eliminating the need for expensive and time-consuming gene annotation. In 1995, Yates *et al.* already presented a software tool to search un-interpreted mass spectra against a six-way translated nucleotide database [105]. However, the true added value of the current proteogenomic studies comes from the use of these peptide identifications in gene finding [99].

Since the initial studies [106, 107], numerous proteogenomic reannotations of genomes have been reported [108–112]. To improve gene annotation, genome sequencing of bacteria and archaea is often performed in parallel with proteogenomic studies: *Mycoplasma mobile* [107], *Thermococcus gammatolerans* [113], *Campylobacter concisus* [114], and *Acholeplasma laidlawii* [115]. The exact translation initiation codon and the signal peptide maturation sites can be deducted from protein N-terminal sequences. Multiple N-terminal sequencing technologies have been applied in proteogenomics studies [109, 116]. In a study on *Deinococcus deserti*, 664 N-terminal peptides were observed, leading to the correction of 63 translation initiation codons in the genome [117]. Similar results have been reported for other bacteria, such as for *Salmonella typhimurium* [118], *Shigella flexneri* [119], *Ruegeria pomeroyi* [120], and *Pseudomonas fluorescens* [121], indicating 10% incorrectly predicted initiations of translation.

The genomic structure of genes differs between prokaryotes and eukaryotes. In prokaryotes, related genes are clustered in operons. They share the same promotor and are transcribed into a single mRNA. Regulation of the individual expression levels occurs at the translation

level. Programmed frame-shifts can produce alternative or truncated proteins and are nearly impossible to predict solely from genomic data. In most eukaryotic genes, the coding regions (exons) are interrupted by non-coding regions (introns). During RNA-splicing, introns are removed and exons are joined and form mature mRNA. In many cases, the splicing process can create multiple unique proteins by combining the exons in different ways, a process called 'alternative splicing'.

Currently, most sequenced genomes are annotated using automated annotation pipelines without manual curation. Gene finding algorithms try to identify the operons in prokaryotes, and the genomic coordinates of exons and the splicing patterns in eukaryotes. They combine evidence from multiple orthogonal sources [122]; *ab initio* gene predictors, large scale transcript sequencing projects and evolutionary conservation among related species. Predicting translation start sites remains a major challenge [123]. Genes that differ in GC composition change coding signals, to the point that the tools have to be 'retrained' for each new genome [124].

The most commonly used start codon is AUG, coding for the amino acid methionine during translation. Several codons are indicators of the translation stop of which TGA, TAA and TAG are the most common. To be able to determine the correct translation stop site, the frame of the terminal exon needs to be correctly predicted first. Both in pro- and eukaryotes, alternative start codons have been reported, e.g. GUG and UUG of the lac-operon in *Escherichia coli*. During proteogenomic studies, many non-standard start sites have been observed and used for reannotation of existing genomes [117, 125].

Despite cDNA evidence, 4000 genes in human do not translate into protein [126]. Currently most gene prediction algorithms are only being trained using transcript sequencing data. The data collected in proteogenomic experiments can be used to improve the performance of self-learning gene annotation algorithms. Since C-terminal peptides are underrepresented in standard proteomics data sets, C-terminal sequencing technologies are very useful to generate data sets of protein termini for proteogenomic purposes.

1.10 Degradomics

Limited proteolysis, also known as protein processing, is one of the many known post-translational modifications. Proteases hydrolyze protein substrate bonds at the protein termini (exoproteases) or inside (endoproteases) their protein substrate targets. Proteolysis can be divided into two general classes: sequential maturation and protein partitioning. During sequential maturation a functional protein part is formed by cleavage of a propeptide. The propeptide is often degraded. During protein partitioning two new protein species are formed

with usually unrelated properties [127]. In either way, new protein termini are formed that are susceptible to further modification, such as acetylation [128]. Given the irreversible nature of proteolysis, proteases require tight regulation. *In vivo*, interactions with natural protease inhibitors, post-translational modifications (including removal of propeptides) and subcellular substrate/protease compartmentalization are important regulatory mechanisms for protease activity. Many physiological processes (e.g. food processing and blood clotting) are controlled by proteolytical processing, often interconnected in a protease web [129]. During a large scale evaluation of thirteen terminomics datasets from *Homo sapiens*, *Mus musculus*, and *Escherichia coli* show that >30% of all N-termini and >10% of all C-termini originate from post-translational proteolytic processing other than classical protein maturation (removal of the initiator methionine, signal peptide and pro-peptide) [130]. More recently, in skin, for 50% of the >2000 identified proteins evidence was found for the presence of stable cleavage products *in vivo* [131]. Infectious microorganisms, viruses and parasites also use proteases as virulence factors. Animal venom often contains proteases to evade host responses and degrade tissue [132]. Many human diseases are associated with uncontrolled proteolytic activities, including cancer [133, 134]. Therefore, proteases represent attractive drug targets.

In 2002 Lopez-Otin defined degradomics as all genomic and proteomic investigations and techniques regarding the genetic, structural and functional identification and characterization of proteases, and their substrates and inhibitors, that are present in an organism. The definition of a degradome is twofold: the complete set of proteases that are expressed at a specific moment, or circumstance, by a cell, tissue or organism. Hence the degradome of a protease is defined as the complete natural substrate repertoire of that enzyme in a cell tissue or organism [135].

In MEROPS (http://merops.sanger.ac.uk), the protease database, all information regarding proteases and their inhibitors is combined [136]. Proteases represent a large enzyme family, with 567 members in humans, but more than half of these proteases have no annotated substrates.

In 2008, Schilling *et al.* presented a procedure called Proteomic Identification of protease Cleavage Sites (PICS) to determine both the prime (C-terminal) and non-prime (N-terminal) side residue preferences of endoproteases, using database searchable proteome-derived primary amine protected peptide libraries. After incubation with a protease, prime side cleavage products are tagged with biotin, isolated and identified by MS. The corresponding non-prime side sequences are derived from databases using bioinformatics [137].

Specific analysis of cleavage sites in denatured peptides does not allow to identify native substrates. The COFRADIC technology developed by the group of Gevaert and the TAILS technology developed by the group of Overall allow to differentially study the effect of a protease

on a proteome. These approaches generate terminal sequence information by 'negative' selection of terminal peptides [97, 138, 139]. The TAILS technology was used to study skin inflammation using protease knock-out and wild type mice [131]. A four-way cross-comparison of healthy and inflamed skin revealed a mechanistic insight on the role of proteolysis in inflammation. Several terminal modifications, in particular pyroglutamate formation and an alternative translation initiation site, were also observed.

Proteolytic cleavage also creates novel substrate C-termini, and analyzing these is complementary to N-terminal proteome studies, given that not all protein N-termini can be identified because of problems related to retention on RP columns and peptide ionization and fragmentation. Further, protein neo-C-termini directly point to substrates of the largely unexplored family of carboxypeptidases. Positional proteomics on protein C-termini remains very challenging, mainly because of the lower reactivity of carboxyl groups versus amino groups [139].

References

[1] Collins, F. S., Lander, E. S., Rogers, J., and Waterston, R. H. (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945, Int. Human Genome Sequencing Consortium.

[2] Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, **35**, D61–D65.

[3] Overall, C. M. (2014) Can proteomics fill the gap between genomics and phenotypes? *Journal of proteomics*, **100**, 1–2.

[4] Rappsilber, J. and Mann, M. (2002) What does it mean to identify a protein in proteomics? *Trends in biochemical sciences*, **27**, 74–78.

[5] Smith, L. M. and Kelleher, N. L. (2013) Proteoform: a single term describing protein complexity. *Nature methods*, **10**, 186–187, Top-Down Proteomics Consortium.

[6] Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561–563.

[7] Anderson, L. and Seilhamer, J. (1997) A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, **18**, 533–537.

[8] Gygi, S., Rochon, Y., Franza, B., and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology*, **19**, 1720–1730.

[9] http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology.

[10] Marques-Bonet, T., et al. (2009) A burst of segmental duplications in the genome of the African great ape ancestor. *Nature*, **458**, 877.

[11] Rogers, J. and Gibbs, R. A. (2014) Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature reviews genetics*, **15**, 347–359.

[12] Mutz, K. O., Heilkenbrinker, A., Lonne, M., Walter, J. G., and Stahl, F. (2013) Transcriptome analysis using next-generation sequencing. *Current opinion in biotechnology*, **24**, 22–30.

[13] McGettigan, P. A. (2013) Transcriptomics in the RNA-seq era. *Current opinion in chemical biology*, **17**, 4–11.

[14] Adamski, J. and Suhre, K. (2013) Metabolomics plafforms for genome wide association studies-linking the genome to the metabolome. *Current opinion in biotechnology*, **24**, 39–47.

[15] Blank, L. M. and Ebert, B. E. (2013) From measurement to implementation of metabolic fluxes. *Current opinion in biotechnology*, **24**, 13–21.

[16] Wilkins, M., et al. (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology (N.Y.)*, **14**, 61–65.

[17] Thadikkaran, L., Siegenthaler, M. A., Crettaz, D., Queloz, P. A., Schneider, P., and Tissot, J. D. (2005) Recent advances in blood-related proteomics. *Proteomics*, **5**, 3019–3034.

[18] Schuchard, M. D., Melm, C. D., Crawford, A. S., Chapman, H. A., Fan, F., Ngowe, C., Ray, K. B., Chen, D. E., and Scott, G. B. I. (2006) One step depletion of twenty high abundance human plasma proteins and concomitant molecular size fractionation of low abundance proteins. *Molecular & cellular proteomics*, **5**, S203–S203.

[19] O'Farrell, P. (1975) High resolution two-dimensional electrophoresis of proteins. *Journal of biological chemistry*, **250**, 4007–4021.

[20] Klose, J. (1999) Large-gel 2-D electrophoresis. *Methods in molecular biology*, **112**, 147–172.

[21] Sitek, B., Sipos, B., Pfeiffer, K., Grzendowski, M., Poschmann, G., Hawranke, E., Koper, K., Kloppel, G., Meyer, H. E., and Stuhler, K. (2008) Establishment of "one-piece" large-gel 2-DE for high-resolution analysis of small amounts of sample using difference gel electrophoresis saturation labelling. *Analytical and bioanalytical chemistry*, **391**, 361–365.

[22] Catherman, A. D., Skinner, O. S., and Kelleher, N. L. (2014) Top Down proteomics: Facts and perspectives. *Biochemical and biophysical research communications*, **445**, 683–693.

[23] Olsen, J., Ong, S., and Mann, M. (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & cellular proteomics*, **3**, 608–614.

[24] Swaney, D. L., Wenger, C. D., and Coon, J. J. (2010) Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *Journal of proteome research*, **9**, 1323–1329.

[25] Link, A., Eng, J., Schieltz, D., Carmack, E., Mize, G., Morris, D., Garvik, B., and Yates, I., J.R. (1999) Direct analysis of protein complexes using mass spectrometry. *Nature biotechnology*, **17**, 676–682.

[26] Washburn, M., Wolters, D., and Yates, J. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature biotechnology*, **19**, 242–247.

[27] Washburn, M., Ulaszek, R., Deciu, C., Schieltz, D., and Yates, I., J.R. (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Analytical chemistry*, **74**, 1650–1657.

[28] Gilar, M., Olivova, P., Daly, A., and Gebler, J. (2005) Orthogonality of separation in two-dimensional liquid chromatography. *Analytical chemistry*, **77**, 6426–6434.

[29] Gilar, M., Olivova, P., Daly, A., and Gebler, J. (2005) Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. *Journal of separation science*, **28**, 1694–1703.

[30] Washburn, M. (2008) Presentation - Principles and applications of shotgun proteomics. *2^{nd} EU-summer school in Proteomic basics*.

[31] Yates, J. R. (2011) A century of mass spectrometry: from atoms to proteomes. *Nature methods*, **8**, 633–637.

[32] Fenn, J., Mann, M., Meng, C., Wong, S., and Whitehouse, C. (1989) Electrospray ionization for mass-spectrometry of large biomolecules. *Science*, **246**, 64–71.

[33] Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T. (1988) Protein and polymer analyses up to m/z 100000 by laser ionization time-of-flight mass spectrometry. *Rapid communications in mass spectrometry*, **2**, 151–153.

[34] Karas, M. and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10000 daltons. *Analytical chemistry*, **60**, 2299–2301.

[35] Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

[36] Hillenkamp, F., Karas, M., Beavis, R. C., and Chait, B. T. (1991) Matrix-assisted laser desorption ionization mass-spectrometry of biopolymers. *Analytical chemistry*, **63**, A1193–A1202.

[37] Beavis, R. and Chait, B. (1989) Cinnamic acid derivatives as matrices for ultraviolet laser desorption mass spectrometry of proteins. *Rapid communications in mass spectrometry*, **3**, 432–435.

[38] Beavis, R. and Chait, B. (1990) High-accuracy molecular mass determination of proteins using matrix-assisted laser desorption mass spectrometry. *Analytical chemistry*, **62**, 1836–1840.

[39] Karas, M. and Kruger, R. (2003) Ion formation in MALDI: The cluster ionization mechanism. *Chemical reviews*, **103**, 427–439.

[40] Knochenmuss, R. (2014) Maldi mechanisms: wavelength and matrix dependence of the coupled photophysical and chemical dynamics model. *Analyst*, **139**, 147–156.

[41] Knochenmuss, R. (2014) Energetics and kinetics of thermal ionization models of maldi. *Journal of The American Society for Mass Spectrometry*, pp. 1–7.

[42] Wiley, W. C. and McLaren, I. H. (1955) Time-of-flight mass-spectrometer with improved resolution. *Review of scientific instruments*, **26**, 1150–1157.

[43] Mamyrin, B. A., Karataev, V. I., Shmikk, D. V., and Zagulin, V. A. (1973) Mass-reflectron a new nonmagnetic time-of-flight high-resolution mass-spectrometer. *Zhurnal Eksperimentalnoi I Teoreticheskoi Fiziki*, **64**, 82–89.

[44] Brown, R. and Lennon, J. (1995) Mass resolution improvement by incorporation of pulsed ion extraction in a matrix-assisted laser desorption/ionization linear time-of-flight mass spectrometer. *Analytical chemistry*, **67**, 1998–2003.

[45] Vestal, M., Juhasz, P., and Martin, S. (1995) Delayed extraction matrix-assisted laser-desorption time-of-flight mass-spectrometry. *Rapid communications in mass spectrometry*, **9**, 1044–1050.

[46] Verentchikov, A. N., Ens, W., and Standing, K. G. (1994) Reflecting time-of-flight mass-spectrometer with an electrospray ion-source and orthogonal extraction. *Analytical chemistry*, **66**, 126–133.

[47] Roepstorff, P. and Fohlman, J. (1984) Proposal for a common nomenclature for sequence ions in mass-spectra of peptides. *Biomedical mass spectrometry*, **11**, 601–601.

[48] Perkins, D., Pappin, D., Creasy, D., and Cottrell, J. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

[49] Eng, J., McCormack, A., and Yates, J. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *Journal of the american society for mass spectrometry*, **5**, 976–989.

[50] Fenyo, D. and Beavis, R. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical chemistry*, **75**, 768–774.

[51] Shilov, I., Seymour, S., Patel, A., Loboda, A., Tang, W., Keating, S., Hunter, C., Nuwaysir, L., and Schaeffer, D. (2007) The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Molecular & cellular proteomics*, **6**, 1638–1655.

[52] Loo, J. A., Quinn, J. P., Ryu, S. I., Henry, K. D., Senko, M. W., and McLafferty, F. W. (1992) High-resolution tandem mass-spectrometry of large biomolecules. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 286–289.

[53] Macek, B., Waanders, L. F., Olsen, J. V., and Mann, M. (2006) Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. *Molecular & cellular proteomics*, **5**, 949–958.

[54] Liu, T., Belov, M. E., Jaitly, N., Qian, W. J., and Smith, R. D. (2007) Accurate mass measurements in proteomics. *Chemical reviews*, **107**, 3621–3653.

[55] Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American chemical society*, **120**, 3265–3266.

[56] Sleno, L. and Volmer, D. A. (2004) Ion activation methods for tandem mass spectrometry. *Journal of mass spectrometry*, **39**, 1091–1112.

[57] Zubarev, R. A., Haselmann, K. F., Budnik, B., Kjeldsen, F., and Jensen, F. (2002) Towards an understanding of the mechanism of electron-capture dissociation: a historical perspective and modern ideas. *European journal of mass spectrometry*, **8**, 337–349.

[58] Leymarie, N., Costello, C. E., and O'Connor, P. B. (2003) Electron capture dissociation initiates a free radical reaction cascade. *Journal of the American chemical society*, **125**, 8949–8958.

[59] Paizs, B. and Suhai, S. (2005) Fragmentation pathways of protonated peptides. *Mass spectrometry reviews*, **24**, 508–548.

[60] Horn, D. M., Ge, Y., and McLafferty, F. W. (2000) Activated ion electron capture dissociation for mass spectral sequencing of larger (42 kDa) proteins. *Analytical chemistry*, **72**, 4778–4784.

[61] Sze, S. K., Ge, Y., Oh, H., and McLafferty, F. W. (2002) Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 1774–1779.

[62] Little, D. P., Speir, J. P., Senko, M. W., O'connor, P. B., and McLafferty, F. W. (1994) Infrared multiphoton dissociation of large multiply-charged ions for biomolecule sequencing. *Analytical chemistry*, **66**, 2809–2815.

[63] Mihalca, R., van der Burgt, Y. E. M., McDonnell, L. A., Duursma, M., Cerjak, I., Heck, A. J. R., and Heeren, R. M. A. (2006) Combined infrared multiphoton dissociation and electron-capture dissociation using co-linear and overlapping beams in Fourier transform ion cyclotron resonance mass spectrometry. *Rapid communications in mass spectrometry*, **20**, 1838–1844.

[64] Tsybin, Y. O., Witt, M., Baykut, G., Kjeldsen, F., and Hakansson, P. (2003) Combined infrared multiphoton dissociation and electron capture dissociation with a hollow electron beam in Fourier transform ion cyclotron resonance mass spectrometry. *Rapid communications in mass spectrometry*, **17**, 1759–1768.

[65] Tsybin, Y. O., He, H., Emmett, M. R., Hendrickson, C. L., and Marshall, A. G. (2007) Ion activation in electron capture dissociation to distinguish between N-terminal and C-terminal productions. *Analytical chemistry*, **79**, 7596–7602.

[66] Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9528–9533.

[67] Suckau, D. and Resemann, A. (2003) T(3)-sequencing: Targeted characterization of the N- and C-termini of undigested proteins by mass spectrometry. *Analytical chemistry*, **75**, 5817–5824.

[68] Suckau, D., Resemann, A., Schuerenberg, M., Hufnagel, P., Franzen, J., and Holle, A. (2003) A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics. *Analytical and bioanalytical chemistry*, **376**, 952–965.

[69] Unlu, M., Morgan, M., and Minden, J. (1997) Difference gel electrophoresis: A single gel method for detecting changes in protein extracts. *Electrophoresis*, **18**, 2071–2077.

[70] Megger, D. A., Bracht, T., Meyer, H. E., and Sitek, B. (2013) Label-free quantification in clinical proteomics. *Biochimica et biophysica acta - Proteins and proteomics*, **1834**, 1581–1590.

[71] Filiou, M. D., Martins-de Souza, D., Guest, P. C., Bahn, S., and Turck, C. W. (2012) To label or not to label: Applications of quantitative proteomics in neuroscience research. *Proteomics*, **12**, 736–747.

[72] Amunugama, R., Jones, R., Ford, M., and Allen, D. (2013) Bottom-up mass spectrometry–based proteomics as an investigative analytical tool for discovery and quantification of proteins in biological samples. *Advances in wound care*, **2**, 549–557.

[73] Abdallah, C., Dumas-Gaudot, E., Renaut, J., and Sergeant, K. (2012) Gel-based and gel-free quantitative proteomics approaches at a glance. *International journal of plant genomics*, **2012**.

[74] Gygi, S., Rist, B., Gerber, S., Turecek, F., Gelb, M., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature biotechnology*, **17**, 994–999.

[75] Ross, P., et al. (2004) Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics*, **3**, 1154–1169.

[76] Oda, Y., Huang, K., Cross, F., Cowburn, D., and Chait, B. (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 6591–6596.

[77] Ong, S., Blagoev, B., Kratchmarova, I., Kristensen, D., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & cellular proteomics*, **1**, 376–386.

[78] Mirgorodskaya, O., Kozmin, Y., Titov, M., Korner, R., Sonksen, C., and Roepstorff, P. (2000) Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using ^{18}O-labeled internal standards. *Rapid communications in mass spectrometry*, **14**, 1226–1232.

[79] Reynolds, K., Yao, X., and Fenselau, C. (2002) Proteolytic ^{18}O labeling for comparative proteomics: evaluation of endoprotease Glu-C as the catalytic agent. *Journal of proteome research*, **1**, 27–33.

[80] Yao, X., Freas, A., Ramirez, J., Demirev, P., and Fenselau, C. (2001) Proteolytic ^{18}O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Analytical chemistry*, **73**, 2836–2842.

[81] Desiderio, D. and Kai, M. (1983) Preparation of stable isotope-incorporated peptide internal standards for field desorption mass spectrometry quantification of peptides in biologic tissue. *Biomedical mass spectrometry*, **10**, 471–479.

[82] Gerber, S., Rush, J., Stemman, O., Kirschner, M., and Gygi, S. (2003) Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 6940–6945.

[83] Kirkpatrick, D., Gerber, S., and Gygi, S. (2005) The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods*, **35**, 265–273.

[84] Liu, H., Sadygov, R., and Yates, I., J.R. (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical chemistry*, **76**, 4193–4201.

[85] Podwojski, K., Eisenacher, M., Kohl, M., Turewicz, M., Meyer, H., Rahnenfuhrer, J., and Stephan, C. (2010) Peek a peak: a glance at statistics for quantitative label-free proteomics. *Expert review of proteomics*, **7**, 249–261.

[86] Stricker, N. L., Christopherson, K. S., Yi, B. A., Schatz, P. J., Raab, R. W., Dawes, G., Bassett, D. E., Bredt, D. S., and Li, M. (1997) PDZ domain of neuronal nitric oxide synthase recognizes novel C-terminal peptide sequences. *Nature biotechnology*, **15**, 336–342.

[87] Doyle, D. A., Lee, A., Lewis, J., Kim, E., Sheng, M., and MacKinnon, R. (1996) Crystal structures of a complexed and peptide-free membrane protein-binding domain: Molecular basis of peptide recognition by PDZ. *Cell*, **85**, 1067–1076.

[88] Chung, J. J., Shikano, S., Hanyu, Y., and Li, M. (2002) Functional diversity of protein C-termini: more than zipcoding? *Trends in cell biology*, **12**, 146–150.

[89] Wilkins, M. R., et al. (1998) Protein identification with N and C-terminal sequence tags in proteome projects. *Journal of molecular biology*, **278**, 599–608.

[90] Chung, J. J., Yang, H. M., and Li, M. (2003) Genome-wide analyses of carboxyl-terminal sequences. *Molecular & cellular proteomics*, **2**, 173–181.

[91] Creasy, D. M. and Cottrell, J. S. (2004) Unimod: Protein modifications for mass spectrometry. *Proteomics*, **4**, 1534–1536.

[92] Sriram, S. M., Kim, B. Y., and Kwon, Y. T. (2011) The N-end rule pathway: emerging functions and molecular principles of substrate recognition. *Nature reviews molecular cell biology*, **12**, 735–747.

[93] Chen, T., Muratore, T. L., Schaner-Tooley, C. E., Shabanowitz, J., Hunt, D. F., and Macara, I. G. (2007) N-terminal α-methylation of RCC1 is necessary for stable chromatin association and normal mitosis. *Nature cell biology*, **9**, 596–U203.

[94] Petkowski, J. J., Tooley, C. E. S., Anderson, L. C., Shumilin, I. A., Balsbaugh, J. L., Shabanowitz, J., Hunt, D. F., Minor, W., and Macara, I. G. (2012) Substrate specificity of mammalian N-terminal α-amino methyltransferase NRMT. *Biochemistry*, **51**, 5942–5950.

[95] Dormeyer, W., Mohammed, S., van Breukelen, B., Krijgsveld, J., and Heck, A. J. R. (2007) Targeted analysis of protein termini. *Journal of proteome research*, **6**, 4634–4645.

[96] Wu, J., Tolstykh, T., Lee, J., Boyd, K., Stock, J. B., and Broach, J. R. (2000) Carboxyl methylation of the phosphoprotein phosphatase 2A catalytic subunit promotes its functional association with regulatory subunits in vivo. *Embo journal*, **19**, 5672–5681.

[97] Lange, P. F. and Overall, C. M. (2013) Protein TAILS: when termini tell tales of proteolysis and function. *Current opinion in chemical biology*, **17**, 73–82.

[98] Lange, P. F., Huesgen, P. F., and Overall, C. M. (2011) TopFIND 2.0 linking protein termini with proteolytic processing and modifications altering protein function. *Nucleic acids research*, p. gkr1025.

[99] Castellana, N. and Bafna, V. (2010) Proteogenomics to discover the full coding content of genomes: A computational perspective. *Journal of proteomics*, **73**, 2124–2135.

[100] Holmes, M. R. and Giddings, M. C. (2008) Using GFS to identify encoding genomic loci from protein mass spectral data. *Current protocols in bioinformatics*, pp. 13–9.

[101] Risk, B. A., Spitzer, W. J., and Giddings, M. C. (2013) Peppy: Proteogenomic Search Software. *Journal of proteome research*, **12**, 3019–3025.

[102] Kumar, D., Yadav, A. K., Kadimi, P. K., Nagaraj, S. H., Grimmond, S. M., and Dash, D. (2013) Proteogenomic analysis of *Bradyrhizobium japonicum* USDA110 using genosuite, an automated multi-algorithmic pipeline. *Molecular & cellular proteomics*, **12**, 3388–3397.

[103] Ferro, M., et al. (2008) Pepline: A software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences. *Journal of proteome research*, **7**, 1873–1883.

[104] Pang, C. N. I., et al. (2014) Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. *Journal of proteome research*, **13**, 84–98.

[105] Yates, J. R., Eng, J. K., and McCormack, A. L. (1995) Mining genome - Correlating tandem mass-spectra of modified and unmodified peptides to sequences in nuleotide data. *Analytical chemistry*, **67**, 3202–3210.

[106] Arthur, J. W. and Wilkins, M. R. (2004) Using proteomics to mine genome sequences. *Journal of proteome research*, **3**, 393–402.

[107] Jaffe, J. D., Berg, H. C., and Church, G. M. (2004) Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics*, **4**, 59–77.

[108] Venter, E., Smith, R. D., and Payne, S. H. (2011) Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *Plos One*, **6**.

[109] Bonissone, S., Gupta, N., Romine, M., Bradshaw, R. A., and Pevzner, P. A. (2013) N-terminal protein processing: a comparative proteogenomic analysis. *Molecular & cellular proteomics*, **12**, 14–28.

[110] Sevinsky, J. R., Cargile, B. J., Bunger, M. K., Meng, F., Yates, N. A., Hendrickson, R. C., and Stephenson, J. L. (2008) Whole genome searching with shotgun proteomic data: Applications for genome annotation. *Journal of proteome research*, **7**, 80–88.

[111] Savidor, A., Donahoo, R. S., Hurtado-Gonzales, O., VerBerkmoes, N. C., Shah, M. B., Lamour, K. H., and McDonald, W. H. (2006) Expressed peptide tags: An additional layer of data for genome annotation. *Journal of proteome research*, **5**, 3048–3058.

[112] Maillet, I., Berndt, P., Malo, C., Rodriguez, S., Brunisholz, R. A., Pragai, Z., Arnold, S., Langen, H., and Wyss, M. (2007) From the genome sequence to the proteome and back: Evaluation of *E-coli* genome annotation with a 2-D gel-based proteomics approach. *Proteomics*, **7**, 1097–1106.

[113] Zivanovic, Y., Armengaud, J., Lagorce, A., Leplat, C., Guerin, P., Dutertre, M., Anthouard, V., Forterre, P., Wincker, P., and Confalonieri, F. (2009) Genome analysis and genome-wide proteomics of *Thermococcus gammatolerans*, the most radioresistant organism known amongst the Archaea. *Genome biology*, **10**.

[114] Deshpande, N. P., Kaakoush, N. O., Mitchell, H., Janitz, K., Raftery, M. J., Li, S. S., and Wilkins, M. R. (2011) Sequencing and validation of the genome of a campylobacter concisus reveals intra-species diversity. *Plos One*, **6**.

[115] Lazarev, V. N., et al. (2011) Complete genome and proteome of *Acholeplasma laidlawii*. *Journal of Bacteriology*, **193**, 4943–4953.

[116] Armengaud, J. (2009) A perfect genome annotation is within reach with the proteomics and genomics alliance. *Current opinion in microbiology*, **12**, 292–300.

[117] Baudet, M., et al. (2010) Proteomics-based refinement of *Deinococcus deserti* genome annotation reveals an unwonted use of non-canonical translation initiation codons. *Molecular & cellular proteomics*, **9**, 415–426.

[118] Ansong, C., et al. (2011) Experimental annotation of post-translational features and translated coding regions in the pathogen *Salmonella typhimurium*. *Bmc genomics*, **12**.

[119] Zhao, L. N., Liu, L. G., Leng, W. C., Wei, C. D., and Jin, Q. (2011) A proteogenomic analysis of *Shigella flexneri* using 2D LC-MALDI TOF/TOF. *Bmc genomics*, **12**.

[120] Christie-Oleza, J. A., Miotello, G., and Armengaud, J. (2012) High-throughput proteogenomics of *Bacillus Ruegeria pomeroyi*: seeding a better genomic annotation for the whole marine *Roseobacter* clade. *Bmc genomics*, **13**.

[121] Kim, W., Silby, M. W., Purvine, S. O., Nicoll, J. S., Hixson, K. K., Monroe, M., Nicora, C. D., Lipton, M. S., and Levy, S. B. (2009) Proteomic detection of non-annotated protein-coding genes in *Pseudomonas fluorescens*. *Plos One*, **4**.

[122] Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M. J., and Clamp, M. (2004) The Ensembl automatic gene annotation system. *Genome research*, **14**, 942–950.

[123] Brent, M. R. (2008) Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature reviews genetics*, **9**, 62–73.

[124] Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology*, **268**, 78–94.

[125] Gupta, N., et al. (2007) Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome research*, **17**, 1362–1377.

[126] Clamp, M., Fry, B., Kamal, M., Xie, X. H., Cuff, J., Lin, M. F., Kellis, M., Lindblad-Toh, K., and Lander, E. S. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19428–19433.

[127] Dean, R. A., Butler, G. S., Hamma-Kourbali, Y., Delbe, J., Brigstock, D. R., Courty, J., and Overall, C. M. (2007) Identification of candidate angiogenic inhibitors processed by matrix metalloproteinase 2 (MMP-2) in cell-based proteomic screens: Disruption of vascular endothelial growth factor (VEGF)/heparin affin regulatory peptide (pleiotrophin) and VEGF/connective tissue growth factor angiogenic inhibitory complexes by MMP-2 proteolysis. *Molecular and cellular biology*, **27**, 8454–8465.

[128] Kleifeld, O., Doucet, A., Keller, U. A. D., Prudova, A., Schilling, O., Kainthan, R. K., Starr, A. E., Foster, L. J., Kizhakkedathu, J. N., and Overall, C. M. (2010) Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nature biotechnology*, **28**, 281–U144.

[129] Doucet, A., Butler, G. S., Rodriguez, D., Prudova, A., and Overall, C. M. (2008) Metadegradomics toward in vivo quantitative degradomics of proteolytic post-translational modifications of the cancer proteome. *Molecular & cellular proteomics*, **7**, 1925–1951.

[130] Lange, P. F. and Overall, C. M. (2011) TopFIND, a knowledgebase linking protein termini with function. *Nature methods*, **8**, 703–704.

[131] auf dem Keller, U., Prudova, A., Eckhard, U., Fingleton, B., and Overall, C. M. (2013) Terminomics reveals MMP2 alters the protease web to increase vessel permeability and complement activity in skin inflammation. *Science signaling*, **6**.

[132] Serrano, S. M. T. (2013) The long road of research on snake venom serine proteinases. *Toxicon*, **62**, 19–26.

[133] Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular Systems Biology*, **7**.

[134] Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011) The quantitative proteome of a human cell line. *Molecular Systems Biology*, **7**.

[135] Lopez-Otin, C. and Overall, C. M. (2002) Protease degradomics: A new challenge for proteomics. *Nature reviews molecular cell biology*, **3**, 509–519.

[136] Rawlings, N. D., Waller, M., Barrett, A. J., and Bateman, A. (2014) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic acids research*, **42**, D503–D509.

[137] Schilling, O. and Overall, C. M. (2008) Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nature biotechnology*, **26**, 685–694.

[138] Van Damme, P., Staes, A., Bronsoms, S., Helsens, K., Colaert, N., Timmerman, E., Aviles, F. X., Vandekerckhove, J., and Gevaert, K. (2010) Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nature methods*, **7**, 512–515.

[139] Plasman, K., Van Damme, P., and Gevaert, K. (2013) Contemporary positional proteomics strategies to study protein processing. *Current opinion in chemical biology*, **17**, 66–72.

Chapter 2

Terminal sequencing technologies

2.1 Introduction

Techniques for sequencing terminal protein regions are a crucial element of terminomics, proteogenomics and degradomics studies. Due to the difference in reactivity of the functional groups, N-terminal sequencing techniques have always been more successful than C-terminal sequencing techniques. Indeed, most methods start with a specific chemical or enzymatic reaction to modify the terminal amino acid that is subsequently released and identified. However, N-terminal labelling is not possible when the N-terminus is 'blocked', e.g. by acetylation, which occurs in as many as 85% eukaryotic cytosolic proteins and which plays an important role in protein stability and protein turnover [1–4]. Furthermore, protein N-termini are highly susceptible to 'trimming' by various aminopeptidases, resulting in proteins that have N-termini missing one, two, or three N-terminal amino acids [5, 6]. This can be problematic when attempting to identify proteolytic cleavage sites, since the N-terminal residue of a neo-N-terminal peptide may not reflect the initial cleavage, but may reflect subsequent trimming events. These drawbacks are less prominent in C-terminal labelling approaches because protein C-termini are rarely blocked compared to N-termini [7] and are not as susceptible to trimming [8].

In the early 50s the first entire protein sequences were determined by chemical sequencing, using paper electrophoresis and spot staining, and later also liquid chromatography and UV detectors, to separate and identify amino acids, mostly in the form of derivatives. With the introduction of soft ionization techniques a new range of mass spectrometry-based sequencing methods was introduced, rapidly followed by the launch of positional proteomics. Below is a detailed overview of these strategies with some examples of N-, but mainly C-terminal sequence determination.

2.2 Chemical sequencing techniques

2.2.1 Sanger sequencing

Frederick Sanger was the first to elucidate the sequence of an entire protein, insulin. After performic acid oxidation of the cystine bridges, the two subunits were characterized [9–11]. The primary structure of these subunits was determined using a mixture of partial acid hydrolysis, 2D-paper chromatography and N-terminal labelling using fluorodinitrobenzene (Sanger-reagent). In 1958 Sanger was awarded the first of his two Nobel prices for the structural work on insulin.

2.2.2 N-terminal Edman degradation

Pehr Edman from Lund University in Sweden introduced the chemistry for the automated N-terminal amino acid sequence determination in 1950 [12]. To submit a protein to Edman sequencing, the protein must first be immobilized on a support. Nowadays, the sample is typically electroblotted from a gel onto a porous polyvinylidene fluoride (PVDF) membrane.

The Edman chemistry cycle consists of three stages; a coupling, cleavage and conversion stage (Figure 2.1). In the coupling step, phenylisothiocyanate (PITC, Edman reagent) is attached under mildly basic conditions (pH 9) to the uncharged α-amino function at the N-terminus of the protein. The newly formed phenylthiocarbamyl (PTC) moiety is cleaved off by adding neat trifluoroacetic acid (TFA). During this cleavage step an unstable cyclic anilinothiazolinone (ATZ)-derivative of the N-terminal amino acid is released from the rest of the unaltered protein. The ATZ-amino acid is extracted from the support and transferred to a conversion flask while the remaining polypeptide can undergo a new coupling step. In the conversion stage the ATZ-amino acid converts into a stable phenylthiohydantoin (PTH)-derivative after treatment with 25% TFA in water. At the end of each cycle of Edman degradation, the PTH-amino acid is separated from reaction by-products and identified by its retention time using C-18 reversed-phase HPLC, combined with UV-absorbance detection. Although not part of the Edman degradation cycle per se, the PTH amino acid analysis step is essential in the protein sequencing process.

The success of the Edman chemistry depends largely on the actual sequence of the peptide of interest. During the reactions, cysteine residues are degraded and therefore, in order to be detected, they need to be modified prior to analysis [13–16]. N-terminally blocked proteins (acetylated, formylated or pyroglutamyl ending) have no free amino group to react with the PITC reagent and have to be 'deblocked' prior to analysis [17, 18].

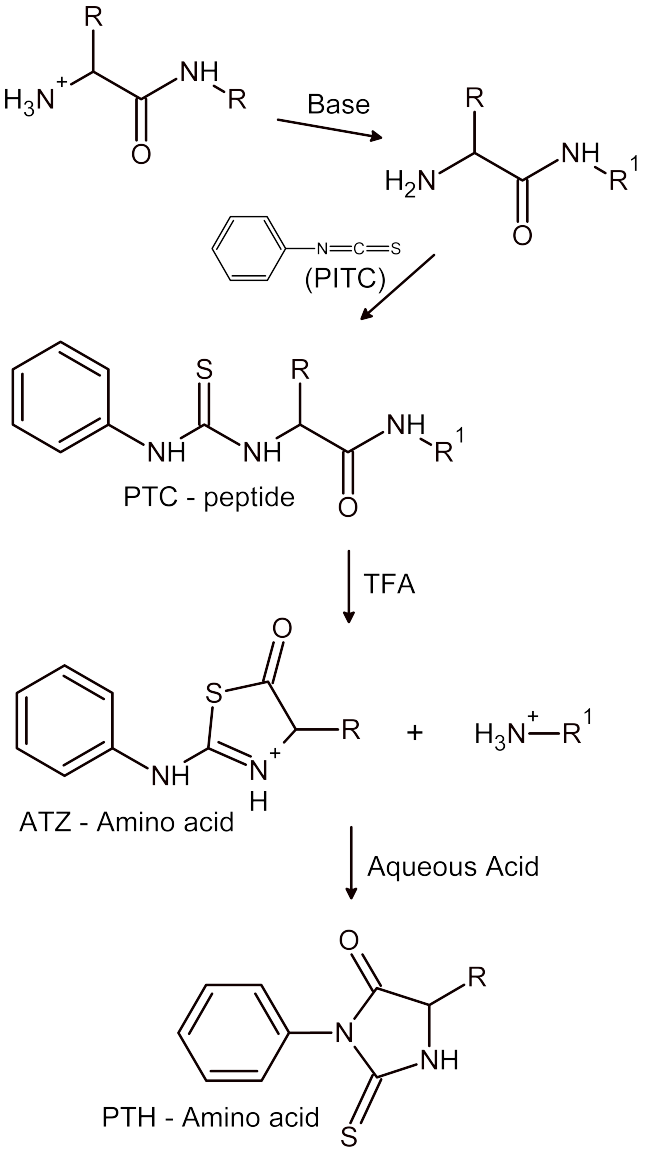


Figure 2.1: The three steps of Edman degradation. Step 1: Coupling of PITC to N-terminal amino acid under basic conditions (trimethylamine) forming a PTC-peptide. Step 2: Cleavage under acidic conditions generates a free amino terminus on the polypeptide and an ATZ-adducted amino acid. Step 3: Conversion of the ATZ-derivative to a PTH-amino acid after treatment with TFA (25%) [19].

The efficiency of the Edman degradation chemistry performed by automated protein sequencers is typically at or above 95%. This allows 30 up to 50 amino acids to be determined in a single run experiment typically starting from 1-50 pmol of separated peptide or protein. A single cycle takes about 30-60 minutes making it possible to run these reactions over multiple days. In order to obtain entire protein sequences, the proteins have to be cleaved and the resulting peptides have to be separated and sequenced individually.

Several by-products are reducing the efficiency. They are formed by reaction of PITC with oxygen, water and dimethylamine, a degradation product of trimethylamine. To prevent the formation of these by-products all reactions are performed under constant argon flow and using 'sequencing-grade' reagents. Besides these by-products masking the PTH separation and interpretation, several other factors limit the reaction efficiency and sensitivity. The cleavage reaction requires a balance between complete cleavage of the ATZ-amino acid from the peptide and unwanted acid cleavage at the other sites along the peptide chain. Each time a non-specific cleavage of the peptide chain occurs, a new N-terminus is formed which can react with PITC, causing the PTH-amino acid background to increase. Also contaminating proteins and free amino acids present in the sample react with PITC, lowering the detection limit.

2.2.3 C-terminal chemical sequencing

The search for a chemical method for stepwise removal of amino acids from the C-terminal end of proteins began early in the 20^{th} century when Schlack and Kumpf applied the method of Johnson and Nicolet, for conversion of acylamino acids to acylthiohydantoins, to a small peptide. The thiohydantoin was then cleaved from the molecule with 1 M sodium hydroxide [20, 21]. Although several methods for chemical sequence analysis have been developed, none have reached the level of efficiency that is achieved using the N-terminal Edman degradation [22].

The nucleophilic α-amino group of the N-terminal amino acid residue permits a facile chemical reaction with PITC under mild reaction conditions, resulting in high repetitive yields (95-99%). The derivatization of the much less chemically reactive carboxylic acid group of the C-terminal amino acid residue proved to be exceedingly more challenging. In several methods the poorly nucleophilic carboxylic group is converted to a more reactive electrophilic functionality. The almost indistinguishable chemical properties of main and side chain carboxyl groups limit the possibilities for selective C-terminal modification. An oxazolone represents one of the very few reactive intermediates that derive solely from the C-terminal carboxyl group of a protein [23].

Acetic anhydride is typically used to activate the carboxyl group forming mixed anhydrides. The activation of the C-terminal carboxyl groups with acetic acid under basic conditions results in the formation of an oxazolone, except for proline [24]. Under strong basic conditions or in the presence of an acylation catalyst (pyridine), the C-terminal oxazolone can react with excess activating reagent to form a less reactive O-acetylated oxazolone. This reaction is known as the first step of the Dakin-West reaction. Lutidin is commonly used as weak base during activation with acetic anhydride [25]. Mixed anhydrides tend to undergo aminolysis unless counteracted by steric hindrance [26]. For this reason, trifluoroacetic anhydride is attractive for the generation of oxazolones with very little possibility of activating the side-chain carboxyl groups to amidation, but well promoting the sequential degradation of peptides [27]. The side

chain carboxyl groups of C-terminal Glu and Asp residues are known to form cyclic anhydrides [28]. Besides carboxyl groups, acetic anhydride also reacts with other functional groups, causing unwanted side reactions. Amino and hydroxyl groups are acetylated, Ser and Thr side chains are dehydrated and form an unsaturated oxazolinone. Phenol can be added to the reaction to convert the oxazolone to a more stable ester and avoid the side-chain acylation of tyrosine hydroxyl groups by oxazolone-activated carboxyl groups. In turn, the ester can effectively react with a nucleophile amine [29].

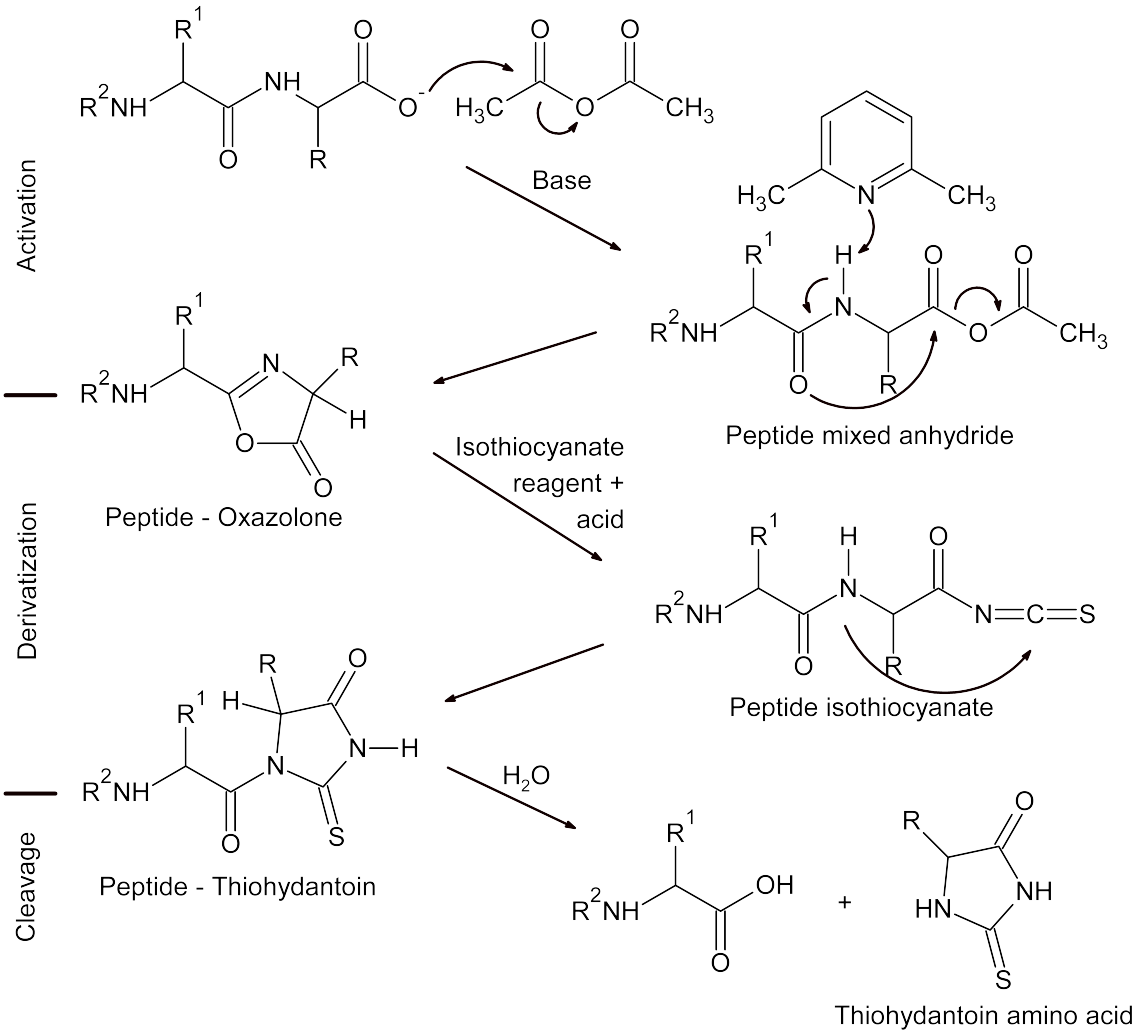


Figure 2.2: Thiocyanate chemistry: Reaction scheme for the sequential C-terminal degradation of peptides using the thiocyanate chemistry. During the activation step, the C-terminal carboxylic group is converted under basic conditions (lutidin) to a reactive electrophilic mixed anhydride or oxazolone. This oxazolone is then derivatized to a thiohydantoin using a thiocyanate reagent under acidic conditions. During the cleavage step, the thiohydantion is cleaved off generating a thiohydantoin derivatized amino acid [30].

The so-called thiocyanate degradation proved to be the most successful chemical C-terminal sequencing method. Traditionally, the steps necessary for one complete cycle of the thiocyanate chemistry have been referred to as the activation, derivatization, and cleavage steps (Figure 2.2). During the activation step the C-terminal carboxylic group is converted to a reactive electrophilic mixed anhydride or oxazolone. This activated carboxylic group is then derivatized to a thiohydantoin using a thiocyanate reagent under acidic conditions. Finally the thiohydantoin derivative of the C-terminal amino acid is specifically cleaved off. Similar to Edman degradation the thiohydantoin derivatives are identified using RPLC analysis. Since thiohydantoin derivatives are more hydrophilic than phenylthiohydantoin derivatives, the separation of the more polar amino acids remains challenging. Applied Biosystems and Hewlett-Packard have automated two variations on the thiocyanate degradation using different activation and cleavage reagents.

DPP-ITC method of Hewlett-Packard

The automated C-terminal sequencer from Hewlett-Packard combines the activation and derivatization step, eliminating the use of acetic anhydride. Diphenyl phosphoroisothiocyanatidate (DPP-ITC) and pyridine are used to convert the C-terminal amino acid of a peptide to a peptidylthiohydantoin. The thiohydantoin is subsequently released through reaction with potassium trimethylsilanolate [24]. The thiohydantoin derivative of the C-terminal amino acid residue is separated and identified by RPLC analysis [31](Figure 2.3).

The use of DPP-ITC to form thiohydantoins was first reported in 1953, but due to slow kinetics (several days) the reaction was abandoned [32]. The addition of pyridine to the reaction mixture allows a complete conversion to a thiohydantoin in less than one hour at 50 °C [33]. After reaction with DPP-ITC, pyridine is delivered in the gas phase resulting in the rearrangement of the pentavalent acylphosphoryl isothiocyanate to the acyl isothiocyanate that rearranges to a thiohydantoin under acidic conditions.

It is claimed that using this chemistry, all natural occurring amino acids can be converted to a thiohydantoin. Proline forms a quaternary ammonium salt containing thiohydantoin. Unlike formation of peptidylthiohydantoins with the other 19 commonly occurring amino acids in which cyclization to a thiohydantoin is concomitant with loss of a proton from the amide nitrogen, proline has no amide proton and, as a result, the newly formed proline thiohydantoin contains an unprotonated ring nitrogen. This cyclic structure, if left unprotonated, will regenerate C-terminal proline during the cleavage reaction. By the addition of acid, the proline thiohydantoin ring is protonated and stabilized. It can be hydrolyzed by the addition of water vapor or sodium trimethylsilanolate to proline thiohydantoin and a shortened peptide. Since

the TFA/water steps have no effect on peptidylthiohydantoins formed from the other 19 amino acids, the additional steps required for proline can readily be integrated into the automated sequencing program [34].

Because PVDF was not stable during the degradation reaction, Zitex (porous teflon) was introduced as support to which the proteins were non-covalently bound. Due to the low initial yield (10-50%) and the low repetitive yield, the sequence analysis of proteins is limited to about 5 C-terminal amino acids starting from nanomolar amounts of purified protein [35].

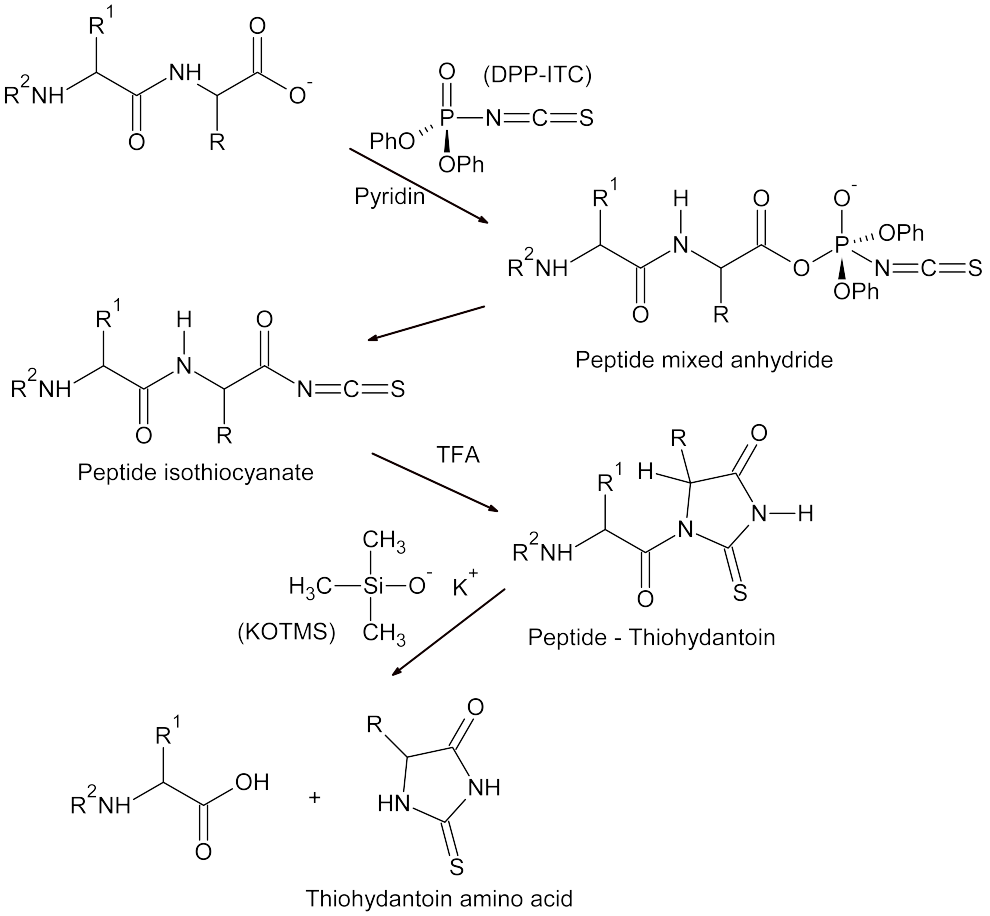


Figure 2.3: Hewlett-Packard chemistry: The detailed reaction scheme of the HP thiohydantoin C-terminal sequence chemistry in which the activation and derivatization step are combined. The chemical coupling reactions with diphenyl phosphorylisothiocyanate (DPP-ITC) generate a mixed anhydride followed by the extrusion of phosphate with ring formation to yield the peptide thiohydantoin. The subsequent chemical cleavage reaction with potassium thimethylsilanolate (KOTMS, a strong nucleophilic base) releases the C-terminal amino acid thiohydantoin derivative from the shortened peptide [31].

Alkylation chemistry method of Applied Biosystems

In 1992 a new method was described for C-terminal degradation, which was implemented, on an automated sequencer by Applied Biosystems (Figure 2.4). The sequencing procedure consists of converting the C-terminal amino acid into a thiohydantoin derivative, followed by transformation of the TH into a good leaving group by alkylation of the sulfur atom with 2-bromomethyl naphthalene. The alkylated thiohydantoin is cleaved mildly and efficiently with thiocyanate, which simultaneously converts the penultimate amino acid to a thiohydantoin [36]. The concomitant cleavage and derivatization implies that the activation must be performed only once. The peptidyl-TH can form a resonance stabilized anion with a low pK_a-value (around 7). Therefore it can be alkylated rapidly in basic conditions.

Certain amino acid side chains such as Cys and Lys residues need to be modified prior to analysis and the protocol was expanded with some additional chemical modification cycles [37–39]. Besides the standard activation cycle, a second activation cycle was added. In the first activation cycle, the carboxylterminus of the proteins is converted to an oxazolinone using acetic anhydride and 2,6-dimethylpyridine. The oxazolone is then converted to a thiohydantoin with tetrabutylammonium thiocyanate in the presence of TFA. During the second activation cycle the side-chain carboxylgroups of Asp and Glu are selectively amidated to improve their detectability, using piperidine thiocyanate in the presence of lutidine [25, 40]. To increase the initial yield, the first cycle is repeated once more after the second activation cycle. Following the activation cycles the proteinyl thiohydantoin is S-alkylated and in the same cycle an acetylation is performed on the hydroxyl groups of Ser and Thr using acetic anhydride and N-methylimidazole as catalyst. This cycle is called the 'OH-capping and sequencing cycle' and is followed by the cleavage of the alkylated thiohydantoin using tetrabutylammonium thiocyanate and TFA vapors. The subsequent cycles only require alkylation and cleavage reactions, and the ATH-products are analyzed using RPLC.

The inability to sequence through proline is one of the main problems in the use of this alkylation chemistry. It was shown that the derivatization of N-acetylproline to proline thiohydantoin is possible with a 100% yield [41].

Both sequencers are discontinued and no longer supported by their manufacturers, although they certainly served a function in the characterization of protein termini of relatively pure, mainly recombinant, proteins. The chemical sequencing techniques lack the sensitivity and throughput to be succesfull on most biological samples.

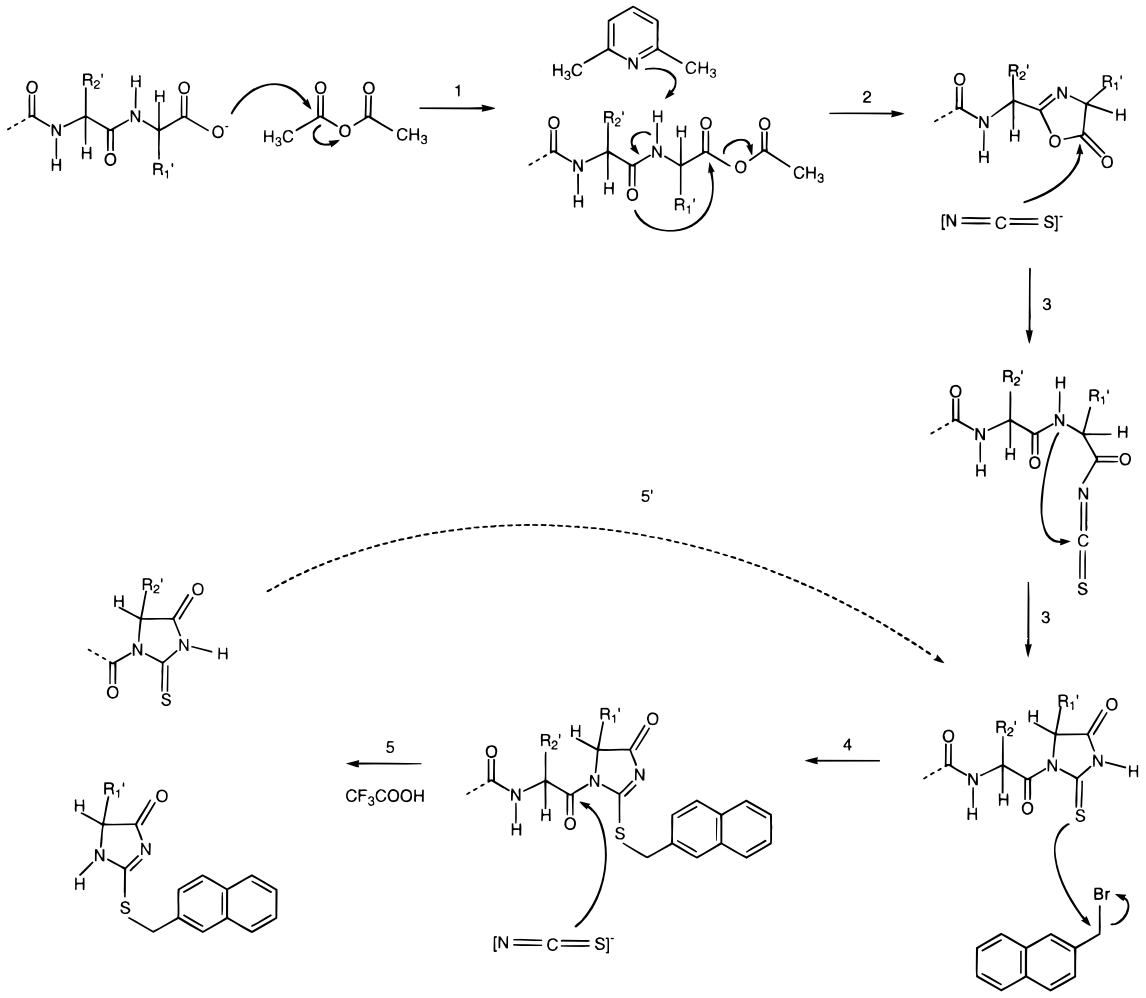


Figure 2.4: Alkylation chemistry: The reaction scheme of the alkylation chemistry as automated by Applied Biosystems. During the cleavage step, the penultimate amino acid is simultaneously converted to form a thiohydantoin. During the activation steps (1 & 2) the C-terminal carboxyl is converted to an oxazolone that in turn is derivatized by the thiocyanate reagent (step 3). During the alkylation step (4) 2-(bromomethyl)naphthalene is coupled to the thiohydantoin under basic conditions forming a good leaving group. Cleavage is induced by thiocyanate reagent under acidic conditions reforming an peptide thiohydantoin [39].

2.3 MS-based sequencing techniques

In the early 90s MS analyzers were introduced as detectors in chemical sequencing setups [34, 42]. When compared to the classic RPLC and UV characterisation of amino acids and their derivatives, MS analyzers offered advantages that included speed, sensitivity and the ability to analyze post-translational modified amino acids [19]. However, they evidently did not overcome the issues relating to the chemical sequencing strategies (increasing background after a certain number of cycles, side reactions) and were never routinely used.

The currently reported terminal sequencing methods are all based on a few strategies to obtain that sequence. In the so-called 'ladder sequencing' approaches a series of sequentially truncated polypeptides is generated. Other techniques selectively target functional groups in order to label, select or enrich the terminal peptide. Recently, LC-MS based methods were reported that enrich terminal peptides using their physicochemical properties. The actual terminal sequence of the protein is obtained in two ways: fragment spectrum analysis of the terminal peptide or interpretation of polypeptide ladders observed in MALDI-MS.

2.3.1 Ladder sequencing techniques

In 1993, Chait *et al.* described the ladder generating chemistry as the controlled generation of a family of sequence-defining peptide fragments from a polypeptide chain, each differing from the next by one amino acid. The resulting sequencing ladder is analyzed by MALDI-MS. Each amino acid is identified by the mass difference between successive peaks and the position in the data set defines the sequence of the original peptide chain [43]. The initial chemical ladder generating techniques were based on the known chemical degradation techniques. Chait used a mixture of 95% PITC and 5% PIC during successive N-terminal degradation cycles without intermediate separation or analysis of the released amino acid derivates. PIC serves as a terminating reagent forming stable phenylcarbamyl peptides and so blocking a small portion of the peptides at each cycle.

The group of Tsugita developed a chemical C-terminal ladder generating method using vapors of fluorinated organic acids or their corresponding anhydrides. An oxazolinone is postulated to be an intermediate [44]. However, the incubation of peptides in acids also results in the cleavage of acid-labile bonds (i.e. at the C-terminal side of aspartic acid and the N-terminal side of serine) [44]. The truncation with perfluoric acid anhydrides suppresses the cleavage of internal peptide bonds and, as such, allows more specific C-terminal sequence analysis [27]. Due to the high reactivity of this reagent and the extremely low reaction temperatures required, the truncation with perfluoric acid anhydride is not easy to perform. Reactive groups can be protected from side reactions through acetylation prior to the truncation reaction. This approach was used to determine the C-terminal sequence of intact proteins, like myoglobin, but requires a lot of pure starting material [45, 46]. Controlled acid hydrolysis was used to generate both N- and C-terminal peptide ladders [47, 48].

Whereas methods for N-terminal ladder sequence analysis are mainly based on chemical ladder-generating procedures [43], proteolytic digestion is the preferred method to generate C-terminal sequence ladders. Carboxypeptidase Y and P are the most common serine exopeptidases that are used for this technique. In the initial carboxypeptidase (CPase) based techniques the

sequence was obtained by analysis of the released amino acids. With the advancement of MS techniques, it became possible to perform direct mass analysis on the peptide fragment ladders formed by incubation of peptides and small proteins with CPase [49, 50]. In 2005, Samyn *et al.* reported a method to selectively form C-terminal peptide sequence ladders from a mixture of CNBr-generated peptides using CPase (Figure 2.5)[51, 52]. During CNBr cleavage, all Met-Xxx peptide bonds are cleaved and all methionine residues are converted to a homoserine-derivative. CPase does not cleave the homoserine lactone ending internal and N-terminal peptides. When a mixture of CNBr generated peptides is incubated with CPase, only the carboxyl ending C-terminal peptide will be degraded. The method has successfully been applied on a proteomic scale and used to study proteolytic processing of procardosin A. CPases cleave off different amino acids at different rates, while certain amino acids are not cleaved off at all [53]. Therefore, to obtain optimal results, the digestion protocol should be specifically optimized for every peptide. A detailed overview of the cyanogen bromide reaction mechanism, enzymatic ladder sequencing techniques and their limitations are given in Chapter 3.

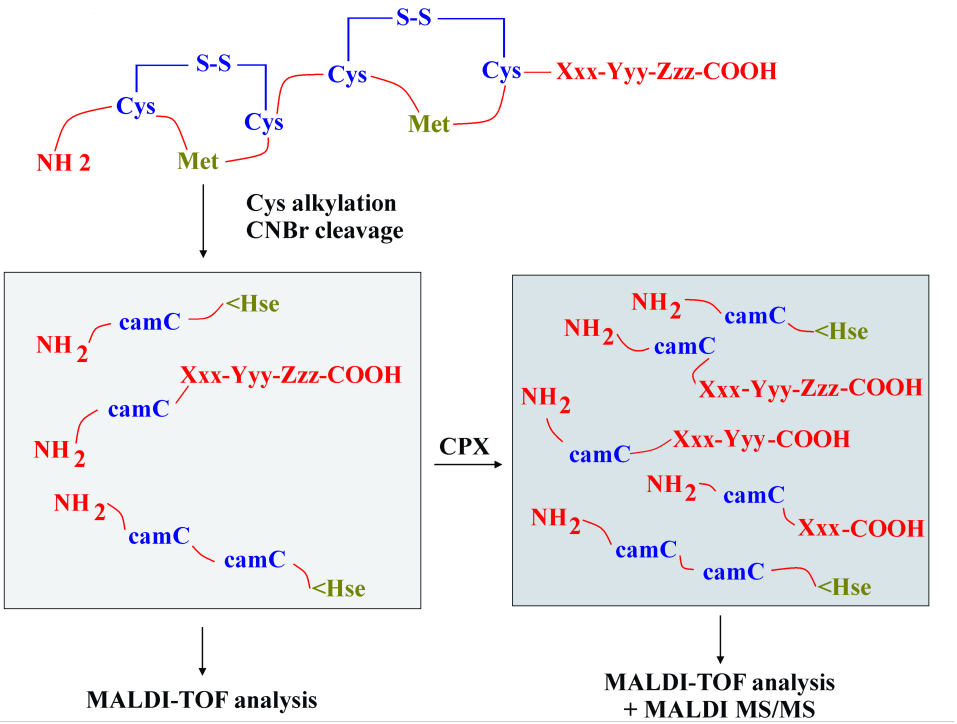


Figure 2.5: Schematic representation of the different steps in the CPase-based C-terminal sequencing method. After reduction, proteins are cleaved by CNBr, generating homoserine lactone ending internal peptides. Only the C-terminal sequence (Xxx-Yyy-Zzz) is accessible for enzymatic degradation by (CPase) and forms a ladder that is analyzed by MALDI mass spectrometry. CPX represents a CPase or mixture of CPases [52].

2.3.2 Selective labelling of functional groups to distinguish terminal peptides

Most of the currently reported techniques, chemically or enzymatically, target specific functional groups in order to label, select or enrich the terminal peptide from the protease digest mixture. The sequence of the terminal peptides is then determined by *de novo* interpretation of tandem mass spectra. The carboxylic and α-amino groups are the obvious groups that are targeted in these approaches. Several reagents and coupling resins have been reported to modify or bind these functional entities. Some reagents discriminate between terminal and side-chain functional groups. Before the different terminal techniques are discussed, the most commonly used modification reagents are brought to mind.

Overview of commonly used reagents for amino and carboxyl group modification

Several reagents are available to modify the primary amino groups in proteins. As discussed above, the Edman reagent (PITC) can be used to modify all primary amino groups. N-hydroxysuccinimide (NHS) and sulfo-N-hydroxysuccinimide (S-NHS) are commonly used activating reagents for carboxyl groups (Figure 2.6). The formed (S)-NHS-esters react with primary amino groups. Because the pK_a of the α-amino group ($pK_a = 8.9$) is considerably lower than that of the ϵ-amino group of lysine ($pK_a = 10.5$), it is possible to selectively label the α-amino groups. Performing the reaction at lower pH ensures that the lysine amines are rarely in the unprotonated state that allows them to react [54].

Under basic conditions (7 N NH_4OH) O-methylisourea selectively modifies all lysines to homoarginine residues [55–59]. Homoarginines have a higher pK_a which has a positive effect on MS detection. During chemical transamination, α-amino groups are selectively converted into a carbonyl group [60, 61]. The selectivity is achieved by the direct participation of the adjacent carbonyl oxygen. P-phenylene diisothiocyanate (DITC) glass originally used as new support during solid phase Edman degradation, can be used to selectively bind free amino groups in a mixture [62].

In peptide synthesis, carboxyl groups are traditionally coupled to amines using carbodiimides as coupling reagents. The diimides react with both side chain and terminal carboxyl groups. The reaction mechanism and conditions are discussed in detail in Chapter 5. Selective labeling of the C-terminal carboxyl group is particularly challenging, since carboxyl groups have a lower reactivity relative to amino groups, and the carboxyl group in aspartate and glutamate residues exhibits a similar reactivity and moreover, is approximately 50 times more abundant in a typical protein [63]. As was discussed in Chapter 2.2.3, an oxazolone represents one of the very few reactive intermediates that derive solely from the C-terminal carboxyl group of

a protein [23]. To limit undesired modification of Tyr side chains and hydrolysis in aqueous solution, Yamaguchi *et al.* converted the oxazolone to a stable active ester by addition of (pentafluoro)phenol. These active esters can in turn effectively react with a nucleophile sich as NH_2-Y [29, 64](Figure 2.7).

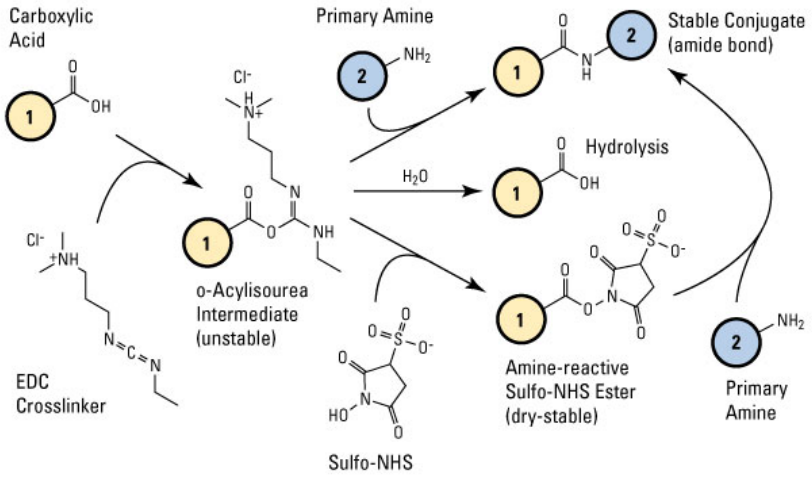


Figure 2.6: NHS and EDC coupling: formation of activated carboxylgroups using carbodiimide and formation of amine reactive sulfo-NHS ester [65].

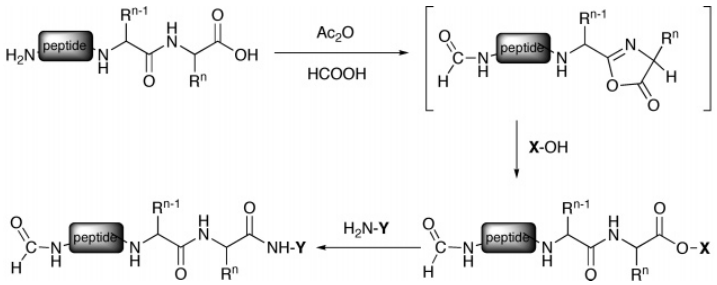


Figure 2.7: Active ester formation via oxazolone: The C-terminal carboxyl group of a peptide reacts with acid anhydride (acetic anhydride in formic acid, in this example) to form an oxazolone via a mixed anhydride. The oxazolone is converted to a more stable active ester (COOX) containing an electron-withdrawing group X. The active ester in turn reacts with a nucleophilic reagent (Y-NH_2) [29].

Labelling with a mass tag

Several methods have been reported to identify the terminal peptide in MS by the presence or absence of certain isotopic distributions or mass differences. By selectively coupling a Br-containing group to the C-terminal carboxyl function, a Br-isotopic signature is observed at the C-terminal peptide in the MS spectrum. A Br-isotopic signature shows a doublet structure

with 2 Da mass difference due to the unique isotopic distribution of bromine (50,7% Br-79; 49,3% Br-81). All y-ions in the fragment spectrum also show the same distribution, simplifying the *de novo* sequence determination [64, 66, 67].

Several isotopic labelling strategies have been reported to allow identifying the C-terminal peptide in a peptide digest mixture. The use of 50% ^{18}O-labeled water during proteolytic digest results in partially heavy labelled internal and N-terminal peptides. Multiple groups have used the technique, although some with mixed success [68, 69]. The labelling efficiency is found to be dependent on both the relative ratio $H_2{}^{16}O$/ $H_2{}^{18}O$ in the digest buffer and on the nature of the peptide formed in a $y=x^2$ relation (y= labelling efficiency and x = relative $H_2{}^{16}O$/ $H_2{}^{18}O$ ratio). Once the free peptide is formed, back-exchange can occur where the peptide-trypsin ester complex is reformed and subsequently hydrolysed. Depending on the ratio of ^{16}O /^{18}O labelled water, this can induce the formation of doubly ^{16}O- or ^{18}O-labelled C-terminal groups. Low pH buffer solutions can also induce back-exchange due to chemical hydrolysis. For trypsin digests typical storage conditions are pH 3-4, low enough to keep the protease inactive and high enough to limit chemical back-exchange [70]. Acylation and esterification using partially d_6-labeled reagents of respectively LysC and GluC generated carboxyl groups have also been used to identify C-terminal peptides [69].

Mass tags can also be introduced by chemical modification of other groups. CNBr cleaves C-terminal of methionine, converting methionine under acidic conditions to homoserine lactone. Incubation of the peptide mixture with acidic methanol results in methanolysis of hsl, adding 32 Da to the peptide. The original carboxyl terminus is converted to a methyl ester, adding 14 Da to the peptide. By comparing the spectra before and after modification, the C-terminus can be identified [71]. Nakazawa *et al.* developed a series of techniques that selectively modify the C-terminal peptide using the oxazolone and the active ester of the oxazolone to specifically introduce chemical groups at the C-terminus. Arginine methyl ester, 2-hydrazino-2-imidazoline and 3-aminopropyltris-(2,4,6-trimethoxyphenyl)phosphonium bromide (TMPP-propylamine) were introduced to enhance the response of the C-terminal peptides during MALDI-ionization [26, 29, 72].

Labelling and enrichment

Since α- and ϵ-amine groups can selectively be modified, multiple techniques have been reported to label and enrich the N-terminal peptide. In 2005, Beynon *et al.* reported a technique to selectively recover N-terminal proteolytic peptides (Figure 2.8 b). All available amino groups are blocked by acetylation using acetic anhydride. Subsequently, the protein is tryptically digested and the newly generated amino groups are biotinylated. The biotinylated internal peptides are removed by recovery onto immobilized avidin or streptavidin, leaving the acetylated

N-terminal peptides in solution. If isotope-labeled acetic anhydride would have been used in the initial labelling, it would have been possible to identify and discriminate between naturally and chemically acetylated peptides [73]. Alternatively, all lysine ϵ-amine groups can be guanidinylated prior to N-terminal amine labelling with a NHS-SS-biotin group. After proteolytic digestion the labelled N-terminal peptides can be positively selected using avidin (Figure 2.8 c) [74, 75]. Variations have been presented where DITC is used to bind the newly formed α-amino groups after selective introduction of a TMPP group at the protein N-terminus (Figure 2.8 e) [76].

In a strategy called N-terminalomics by Chemical Labeling of the α-Amine of Proteins (N-CLAP), the Edman reagent (PITC) is used to label all amino groups. Reaction with TFA selectively unblocks the N-terminus by removal of the N-terminal amino acid. Labelling of the new amino group with Sulfo NHS-SS-Biotin allows affinity enrichment and elution with a reducing agent [6]. In a technique called Dimethyl Isotope-Coded Affinity Selection (DICAS) all amino groups are reductively aminated with formaldehyde, after trypsin digest the newly formed amino groups are depleted using POROS-AL [77, 78] (Figure 2.8 d).

Similar methodologies have been developed where C-terminal peptides are isolated after α-amino group modification combined with LysC digests. DITC glass was used to couple the ϵ-amine groups and, together with the α-amino group labelling, MALDI response enhancing groups were coupled to the C-terminal peptides [72, 79, 80]. The same group recently reported the use of an Arginine-capturing material (m-aminophenylboronic acid-agarose) that, in combination with ArgC, results in isolated C-terminal peptides [81].

Recently, a number of methods have been reported that combine mass and enrichment tagging in multiple derivatization steps. The sequential derivatization reactions are performed while the peptides are bound to the solid phase extraction support (C18 ZipTip). The reagents are sequentially exchanged directly on the resin bed, eliminating intermittent sample purification [82, 83].

Similar to ladder sequencing, enzymes can be used to selectively target peptide backbone carboxyl functions. Anhydrotrypsin is a catalytically inert derivative of trypsin that binds peptides containing lysine or arginine residues at their C-terminus without cleaving them. Immobilized anhydrotrypsin can be used to bind all internal and N-terminal peptides generated by Lys-C digest. The C-terminal peptides remain in solution and are analyzed by LC-MS [84]. Under certain reaction conditions carboxypeptidase Y (CPY), besides proteolysis, exhibits also transpeptidase activity, in which an exogenous nucleophile, such as an amino acid, is added to a protein. In the Profiling Protein C-Termini by Enzymatic Labeling (ProC-TEL) CPY was used to couple a biotin group to the C-terminal carboxylgroup. After trypsin digestion

the C-terminal peptide could be targeted by avidin agarose [8]. A similar positive selection N-terminal sequencing technique exists. A genetically constructed variant of the subtilisin protease, subtiligase, is used to selectively biotinylate unblocked protein α-amines with absolute selectivity over ϵ-amines of lysine side chains. Using 50 mg of starting material, 333 cleavage sites in 292 proteins were characterized when the technique was used to study substrates targeted by caspase-like proteolysis in Jurkat cells following induction of apoptosis [85].

Most of the previously described techniques use a negative selection step to identify the C-terminal peptide. When a negative selection is made, the non-terminal peptides are the ones that are targeted during the enrichment or MS step. The peptides that are not targeted are the ones of interest. Because terminal peptides are far less abundant in a protein digest, the enrichment steps need to be very performant to result in a sample that exclusively contains terminal peptides. Some enrichment techniques using negative selection include an extra modification step to incorporate a mass tag to give a second, positive identification argument during MS/MS analysis. ProC-TEL was the first positive selection technique using enzymatic labeling. The mass tag generated by TMPP serves to exclude false positive identifications. The performance of most techniques depends largely on the length of the generated terminal peptide. MALDI-based techniques cannot detect molecules below 1 kDa and above 5 kDa. Trypsin is used in most techniques, but due to the relatively high amount of charged amino acids in termini, mainly small terminal peptides are generated. Most techniques that use proteases can also not be used when the terminal amino acid is the target of the modification. Proteins ending on Arg and Lys also get ^{18}O-labeled C-terminally by trypsin, Arg-ending peptides also get linked to the m-aminophenylboronic acid agarose matrix.

2.4 LC-MS based proteome wide techniques for terminal sequencing

Similar to standard bottom-up strategies, all terminal sequencing strategies using MS as detector suffer from sample complexity. By introducing an LC separation prior to MS analysis, more complex samples can be analyzed. Several terminal sequencing techniques have been reported to use chromatography as part of the terminal peptide enrichment and identification procedure.

Being able to identify 263 annotated and 87 unpredicted acetylated N-termini and 168 annotated and 193 unpredicted C-termini in the crude membrane fraction of human embryonic carcinoma cells, Heck *et al.* showed that a SCX separation results in the selective enrichment of acetylated N-terminal tryptic peptides and C-terminal peptides [86]. In general, ignoring missed cleavages, tryptic peptides have a lysine or arginine residue at the C-terminus and one free amine group at the N-terminus. At low pH, most acidic residues will be neutral and thus the bulk of the peptides generated will possess two positive charges and will co-elute under SCX chromatography.

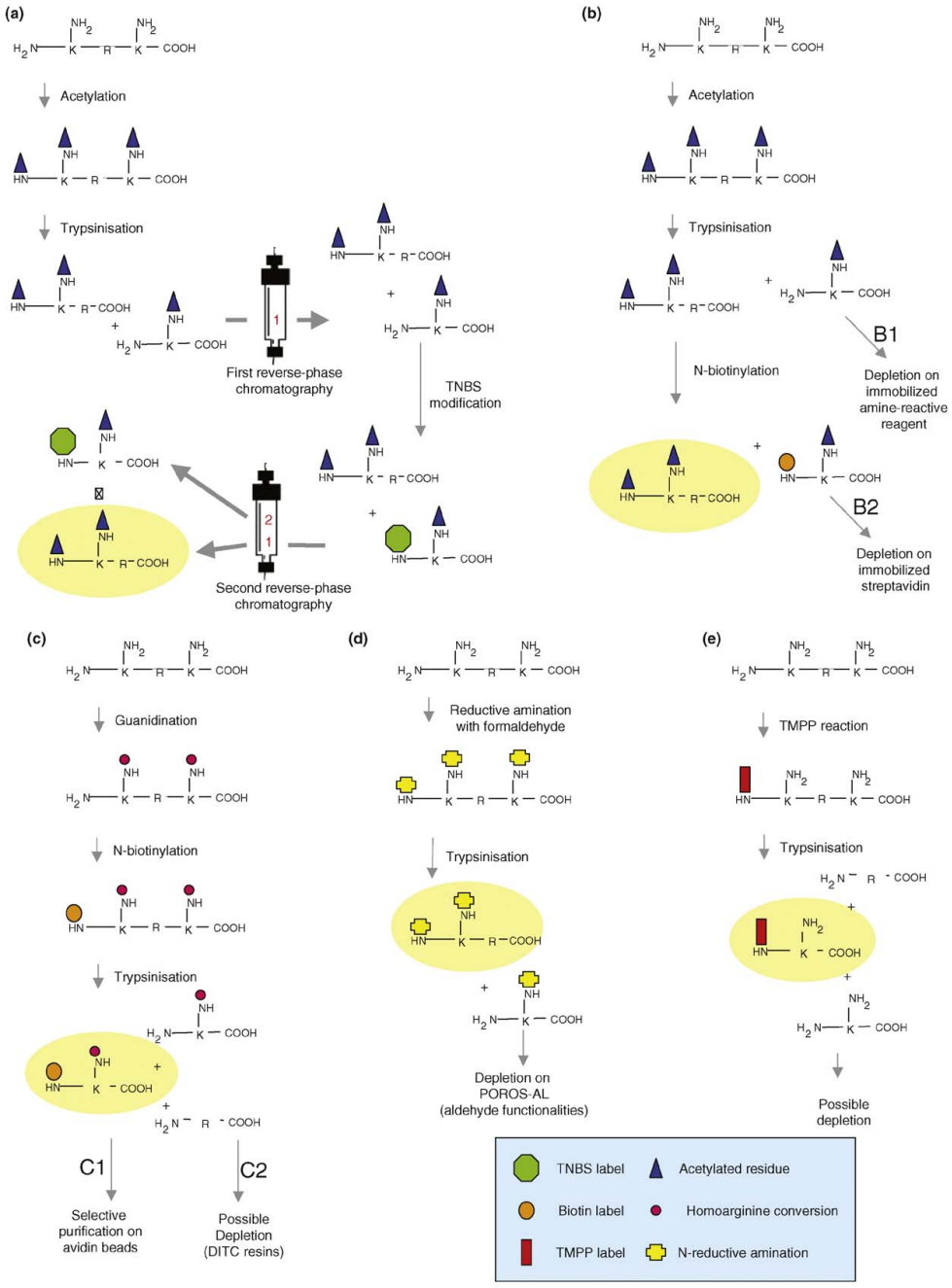


Figure 2.8: Five approaches to specifically enrich N-terminal peptides are shown: COFRADIC (a), acetylation and biotinylation (b), guanidinylation and biotinylation (c), DICAS (d) and TMPP modification (e). N-terminal peptides analyzed by MS/MS are shown with a yellow background. The first step common to all these strategies, consisting in reduction and alkylation of cysteines is not shown. Depending on the strategy used, naturally blocked N-terminal peptides are depleted, enriched and detected or directly analyzed. Alternatives (B1 and B2 and C1 and C2) have been proposed to deplete internal peptides or enrich labelled peptides. In the COFRADIC method (a), eluted fractions 1 and 2 (more retained) from both reverse-phase chromatographies are schematized [87].

Peptides formed from the protein C-terminus will lack a terminal lysine or arginine residue and thus will contain a single positive charge. Peptides corresponding to the blocked (i.e. N-acetylated) N-terminal peptide will also possess a single positive charge. C-terminal peptides and blocked N-terminal peptides can thus be distinguished from other tryptic peptides on basis of the difference in positive charge [88]. Gorman *et al.* first reported the isolation of terminal peptides using cation exchange chromatography in 1993, but the technique was mainly used to enrich tryptic posphopeptides [89, 90].

The group of Gevaert *et al.* set up one of the most successful proteomics platforms using their COFRADIC technology. The approach is based on two papers describing the use of diagonal paper electrophoresis and diagonal chromatography [91, 92]. Diagonal chromatography essentially consists of two identical peptide separation steps with a chemical or enzymatic reaction applied to the fractions in between. This reaction specifically alters the side chain of a specific type of amino acids, thereby changing the chromatographic properties of the peptides holding the targeted (altered) amino acids. The diagonal electrophoresis in combination with Edman sequencing was used to identify the C-terminal peptide sequence of several test proteins and of a *Bacillus* toxin [93, 94].

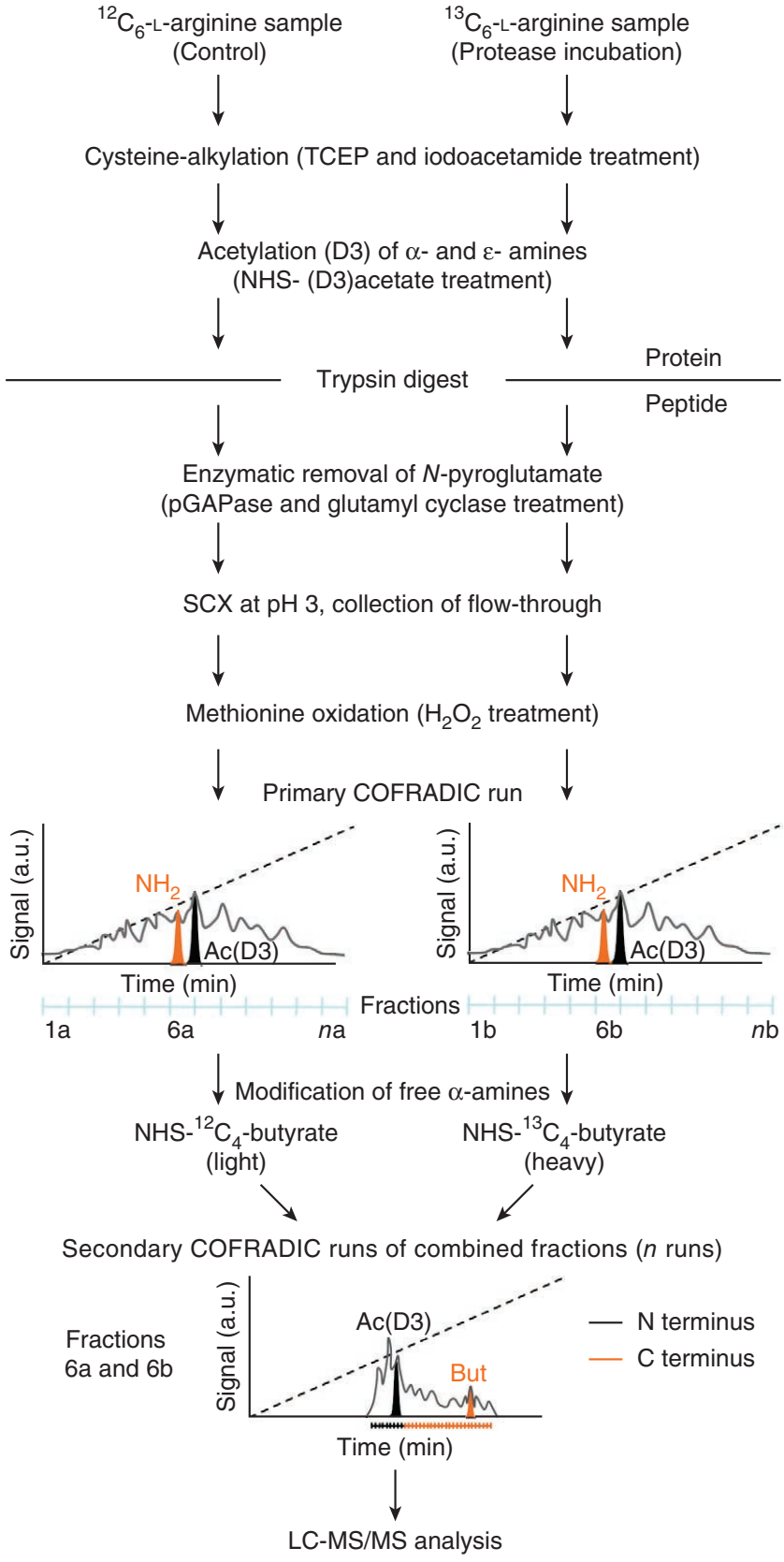
Gevaert *et al.* increased the throughput of the technique by combining several altered primary fractions prior to secondary separations, hence the acronym Combined FRActional Diagonal Chromatography [95]. Changing the reaction selects different classes of peptides. Similar to Isotope coded affinity tag (ICAT) (see Table 1.1), the initial application of the technique focussed on reducing the sample complexity by affinity isolation of a specific class of peptides prior to analysis, the general idea being that when mass spectrometers are not flooded by peptides, more peptide ions are finally fragmented and identified [96]. In the initial COFRADIC setup, controlled oxidation of the side chain of methionine to its sulfoxide counterpart was used to introduce a hydrophillic shift during the second chromatographic separation. Besides the initial Met and Cys labelling techniques [95, 97], different procedures were later developed to study different PTMs: phosphorylations [98], N-glycosylated peptides [99], sialylated N-glycopeptides [100], as well as peptides holding conjugated ATP-derivatives [101] and, more recently, protein ubiquitination identifying over 7500 endogenous ubiquitination sites in human Jurkat cell lysates [102].

In 2003 the COFRADIC N-terminal peptide sorting protocol was presented [5]. After initial alkylation of cysteines and acetylation of all free amino groups, the proteins were tryptically digested (only C-terminal of Arg) and chromatographically separated. Twelve peptide fractions were collected of this primary run and treated with 2,4,6-trinitrobenzenesulfonic acid (TNBS), blocking all free amines. When the TNBS-treated fractions were individually rerun on the same

column under identical conditions, the internal (trinitrophenyl)peptides had shifted to later elution times (hydrophobic shift), whereas the unaltered N-terminal peptides eluted within the same time interval and were collected in a number of secondary fractions. Proline and pyroglutamate residues do not react with TNBS and are detected as false positives, together with the N-terminal blocked peptides. Two additional steps were therefore added to the protocol to eliminate these false positives, a SCX step and an enzymatic step liberating pyro/-glu/-ta/-myl peptides. The SCX step reduces the complexity of the analyte mixture by enriching N-terminal peptides and depleting α-amino free internal peptides, as well as proline-starting peptides. Combined glutamine cyclotransferase (Qcyclase) and pyroglutamyl aminopeptidase (pGAPase) incubation removes all pyroglutamyl residues. Using the improved protocol, 95% of the COFRADIC sorted peptides were found to be α-amino acetylated [103].

In 2010 a slightly modified protocol was applied in a degradomics study to screen for human protein substrates of granzyme B and carboxypeptidase A4 in human cell lyates (Figure 2.9). Cell lysates, metabolically labeled with heavy amino acids were incubated with a protease, or left untreated. Both samples were then treated as mentioned above, except for the use of trideutero-acetylation to differentiate between *in vivo* and *in vitro* acetylation, and the use of heavy and light NHS-C4-butyrate instead of TNBS (Figure 2.10). The two samples were pooled prior to the secondary COFRADIC runs. Since the flow through of SCX at pH 3 should only contain α-amino blocked peptides and C-terminal peptides, all peptides that are hydrophobically shifted after butyrate labelling can be identified as C-terminal peptides. Using this approach 965 protein C-termini, 334 neo-C-termini resulting from granzyme B processing and 16 neo-C-termini resulting from carboxypeptidase A4 processing were identified, besides 1621 protein N-termini [104]. Important to note is the complexity of the protocol and the data sets generated: analysis of one sample results in a total of 52 RPLC runs (Figure 2.11). To improve the peak capacity and protein identifications, multiple columns have been coupled and ran at elevated temperatures, resulting in long runs and GC like resolution [105].

Figure 2.9: Outline of the COFRADIC based positional proteomics procedure using SILAC labeled cell lysates. After protein S-alkylation, reduction and trideutero-acetylation of primary amines, the proteome is digested with trypsin, which now only cleaves after arginine residues. Following pyroglutamate residue removal amino-blocked and C-terminal peptides are enriched during SCX separation. Two parallel RP-HPLC runs with equal amounts of peptide material from the control and the protease-treated proteome are performed. Fractionated peptides containing α-amines (C-terminal peptides) are labeled with different isotopic variants of the amine-reactive NHS-butyrate. Corresponding fractions in time are pooled and reseparated by RP-HPLC, upon which butyrylated C-terminal peptides segregate from the N-terminal peptides. In this way, N-terminal peptides, C-terminal peptides or both are selected for LC-MS/MS analysis. Ac(D3) indicates *in vivo* free, thus *in vitro* trideutero-acetylated peptides, But indicates butyrylated C-terminal peptides [104].


$^{12}C_6$-L-arginine sample (Control)
$^{13}C_6$-L-arginine sample (Protease incubation)
Cysteine-alkylation (TCEP and iodoacetamide treatment)
Acetylation (D3) of α- and ε- amines (NHS- (D3)acetate treatment)
Trypsin digest
Protein
Peptide
Enzymatic removal of N-pyroglutamate (pGAPase and glutamyl cyclase treatment)
SCX at pH 3, collection of flow-through
Methionine oxidation (H_2O_2 treatment)
Primary COFRADIC run
Signal (a.u.)
NH2
Ac(D3)
Time (min)
Fractions
1a
6a
na
1b
6b
nb
Modification of free α-amines
NHS-$^{12}C_4$-butyrate (light)
NHS-$^{13}C_4$-butyrate (heavy)
Secondary COFRADIC runs of combined fractions (n runs)
Fractions 6a and 6b
Ac(D3)
But
N terminus
C terminus
LC-MS/MS analysis

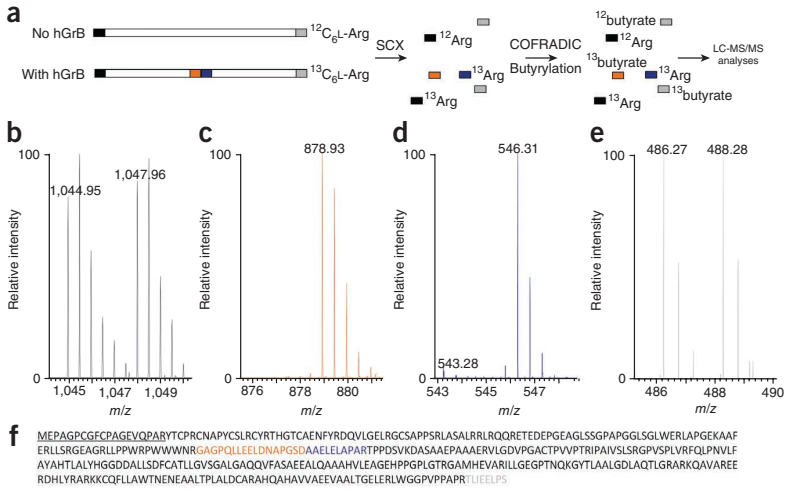


Figure 2.10: Overview of isotopic labels used in COFRADIC for positional proteomics of granzyme B (hGrB) treated and untread samples. N-terminal, C-terminal, neoN-terminal and neoC-terminal peptides and spectra are presented in black, gray, blue and orange respectively. The database annotated N- and C-termini are observed as isotopic doublets of 6 and 4 Da difference respectively (b and e), whereas the neoC terminus and neoN terminus proteolytic signature peptides are identified as singletons (c, d) [104].

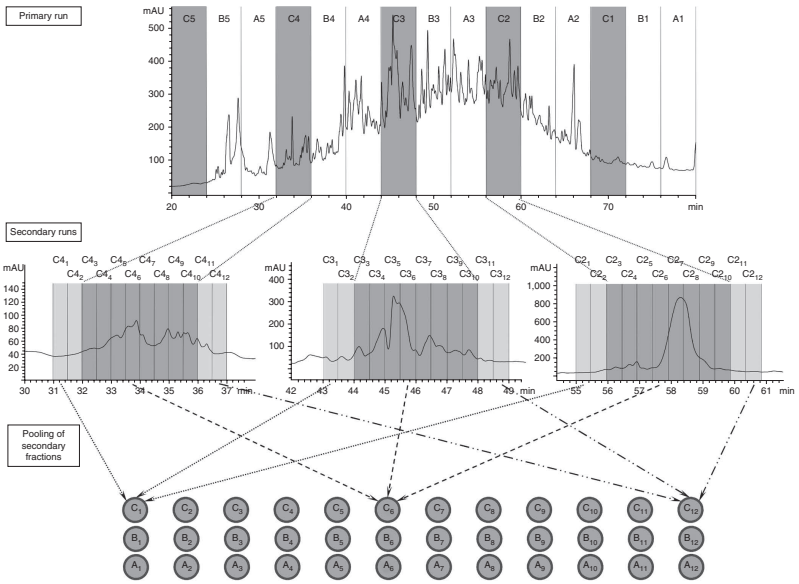


Figure 2.11: Examples of primary and secondary COFRADIC RP-HPLC runs. During the primary run, 15 fractions of 4 minutes each are generated. After modification, these fractions are rerun. During every secondary run, peptides eluting 1 minute before and 1 minute after the primary collection interval are also collected to correct for possible peak broadening (light grey). That way 12 fractions of 30 seconds each are collected during every secondary separation. Multiple fractions of different secondary separations are pooled as illustrated, resulting in a total of 36 samples to be analysed on LC-MS/MS [106].

In 2010, Overall *et al.* presented a C-terminomics procedure, a strategy for the specific isolation and analysis of C-terminal peptides from complex proteomes [107] (Figure 2.12). After reduction and alkylation of protein thiol groups, all amino groups are reductively methylated. All carboxylgroups are then protected by carbodiimide-mediated and N-hydroxysuccinimide-assisted condensation with ethanolamine. Tryptic digestion yields a C-terminal peptide with a protected carboxyl group, which differs from the internal and N-terminal peptides that have free C-termini. These free C-termini are then coupled by 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) mediated condensation to the primary amines of a high-molecular weight linear polyallylamine polymer. The blocked original C-terminal peptides remain unbound and are readily separated from the polymer by ultrafiltration and analyzed by LC-MS/MS. The ethanolamine label serves as positive identification criterium after the negative selection step. Prior to the coupling step the newly generated N-termini of tryptic and C-terminal peptides are protected via a second reductive methylation step. This prevents cross-reactivity, peptide concatamerization or cyclization.

To quantify cleavage events specific to the protease of interest and to distinguish these from proteolysis products present in an untreated sample, the group also introduced two stable isotope label-based techniques to determine the relative abundances of peptides in protease-treated and control samples, Terminal Amine Isotopic Labeling of Substrates (TAILS) and C-Terminal Amine-based Isotope Labeling of Substrates (C-TAILS) [107, 108]. In TAILS, half of the staring material is treated with a protease to form both a control and a protease-treated sample (Figure 2.13). In a reductive methylation step using light and heavy formaldehyde, all amines are labeled prior to enzymatic digestion. All newly formed internal peptides are captured using an amine-reactive polymer and the blocked N-terminal peptides are recovered by centrifugation in a negative selection step and can be analyzed by LS-MS/MS. Similar to TAILS, a stable isotope is incorporated in the tryptic peptides during the second reductive methylation step of C-TAILS using the heavy isotopic form of formaldehyde [107]. An additional SCX step has also been implemented to pre-enrich the acetylated N-terminal peptides prior to analysis, and multiple proteases (trypsin, Lys-C and Lys-N) have been combined to cover a larger part of the N-terminal proteome [109]. As part of the chromosome-centric Human Proteome Project, the TAILS protocol has recently been applied to the human erythrocyte proteome identifying 1369 natural and 1234 neo-N-termini [110].

Overall *et al.* use in-house generated polymers to couple the peptides. To bind all free amines in TAILS a dendritic hyperbranched polyglycerol aldehyde polymer of 100-600 kDa is used and a 56 kDa linear carboxyl reactive polyallylamine polymer is used in C-TAILS. These polymers have a binding capacity of 2.5 mg peptide/mg of polymer, a more than 10 fold improvement in capacity over previous non-specific reactive resins. The improved sample recovery allows to reduce the sample amount to around 100 μg [108].

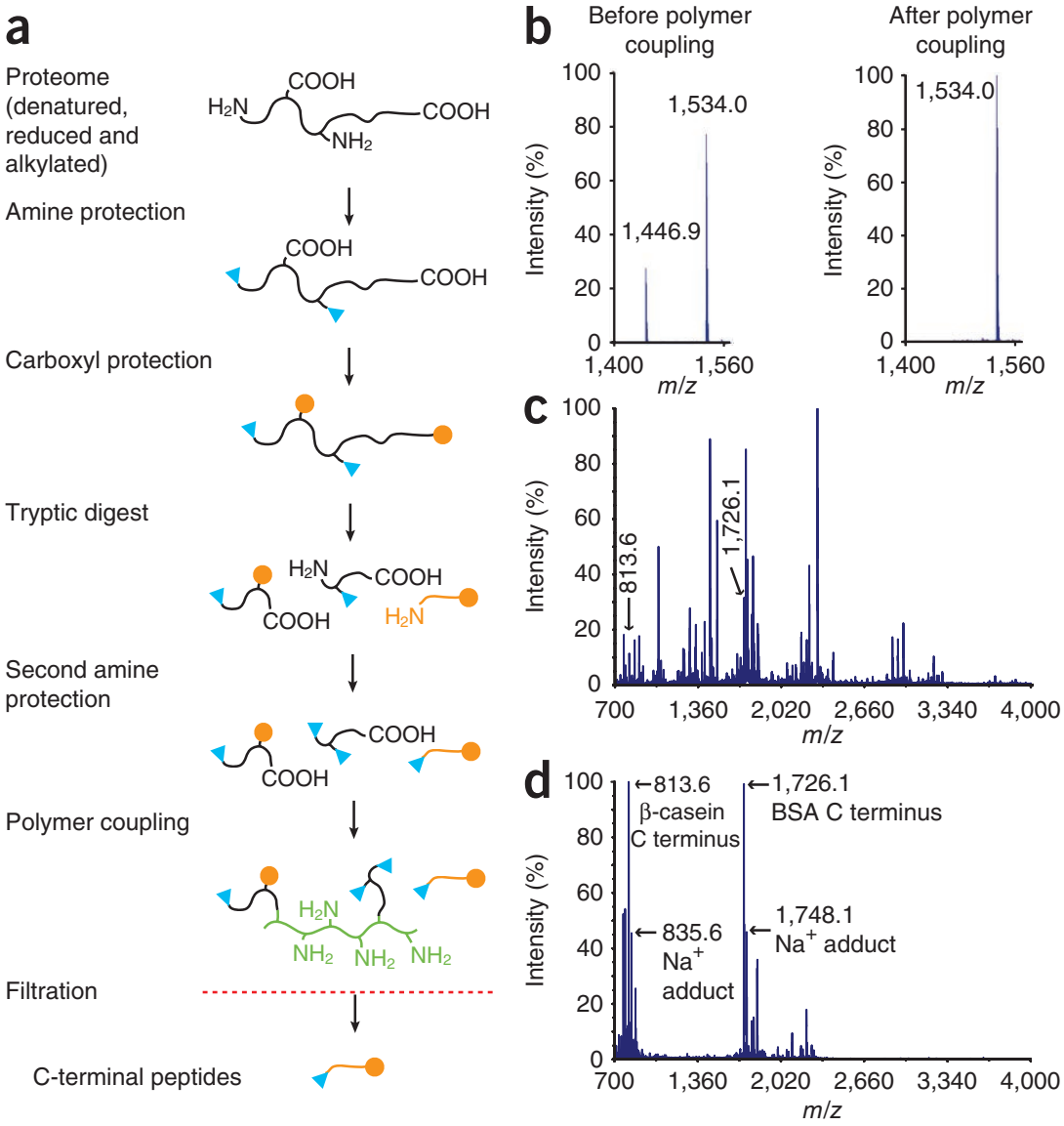


Figure 2.12: C-terminomics workflow. Panel A: overview of the protocol as described in the text. Panel B: Polymer-based enrichment of a peptide with and without protected carboxyl group before and after polymer coupling. Panel C and D: MS spectra of sample of C-terminal peptides of β-casein (812.6 Da) and BSA (1726.1 Da) before (C) and after (D) enrichment [107].

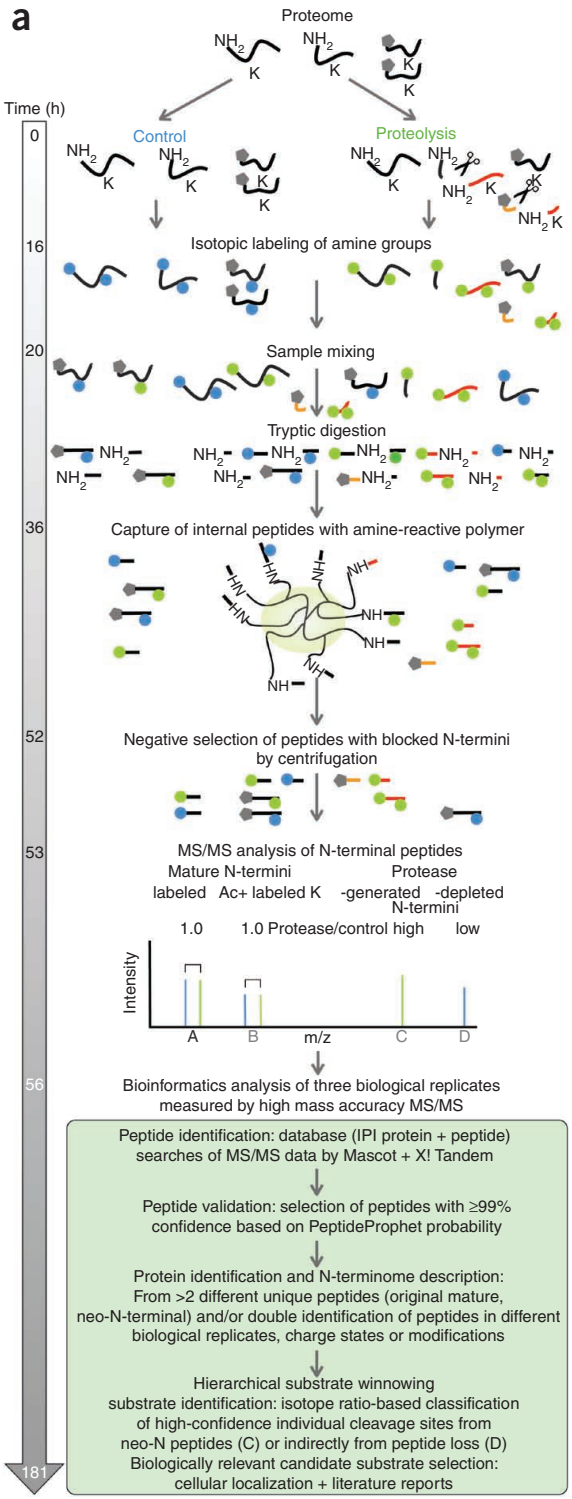


Figure 2.13: Schematic overview of the TAILS technology. Protease treated and control sample are differentially labeled using light and heavy formaldehyde during a reductive methylation step, blue and green circles respectively. The grey pentagons represent N-terminally blocked peptides. An amine-reactive polymer is used to capture all free amino group containing tryptic peptides, resulting in the enrichment of neo and native N-terminal peptides[108].

Recently, a method for absolute quantification of proteolytical processing has been presented that can be applied on a proteome-wide scale. The strategy can be used to monitor both the proteolytic activation of some proteases and the proteolytic cleavage of substrates [111]. Two peptides need to be characterized and chemically synthesized prior to analysis; one tryptic peptide containing the cleavage site of the substrate (Proteolytic Signature Peptide (PSP)) and one peptide common to both the processed and the mature form of the protein substrate (STandard of Expressed Protein (STEP)). The peptides of a digested proteome are labeled with Isobaric tags for relative and absolute quantitation (iTRAQ) and the two standard peptides are labeled with a different form of the iTRAQ reagent. Mixing of the two samples allows selecting one peptide mass with 2 different iTRAQ labels for MS/MS. The ratio of the precursor ions 113 Da and 114 Da allows to determine the ratio of test peptide and proteolytically digested PSP, hence determining the activity of the studied protease on the substrate (Figure 2.14).

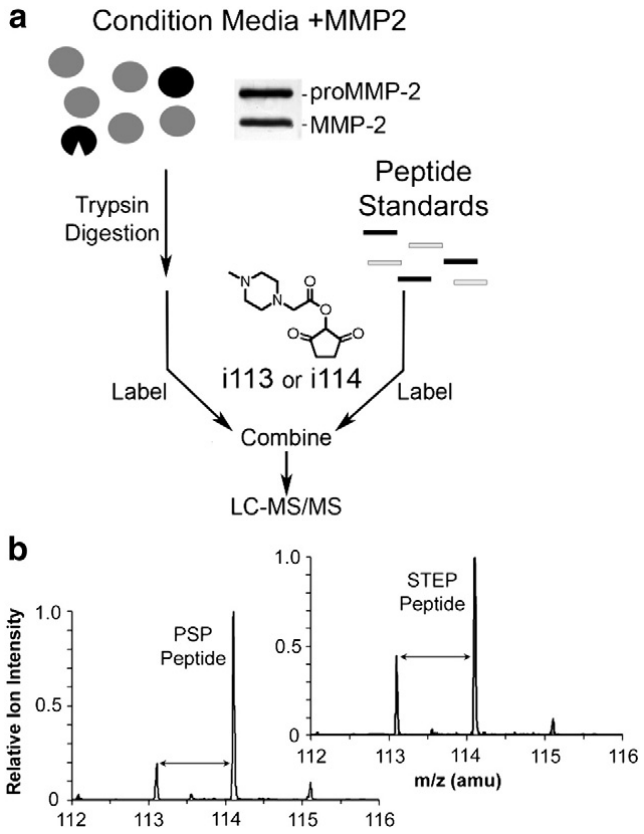


Figure 2.14: Quantitative analysis of proteolysis in a complex mixture. Panel A: Two proteases proMMP-2 and MMP-2 were added to the secreted proteins isolated from conditioned media of Mmp2 -/- embryonic fibroblasts. The sample was digested with trypsin and labeled with iTRAQ-113 and combined with PSP and STEP peptide standards labeled with iTRAQ 114. Panel B: Reporter ions upon MS/MS analysis of the intact PSP and STEP peptides. The relative 113/114 reporter ion ratios and the known concentrations of both peptides allows to determine the concentration of the cleaved substrate. [111].

References

[1] Polevoda, B. and Sherman, F. (2000) N-α-terminal acetylation of eukaryotic proteins. *Journal of biological chemistry*, **275**, 36479–36482.

[2] Driessen, H. P. C., Dejong, W. W., Tesser, G. I., and Bloemendal, H. (1985) The mechanism of N-terminal acetylation of proteins. *Crc critical reviews in biochemistry*, **18**, 281–325.

[3] Varshavsky, A. (2011) The N-end rule pathway and regulation by proteolysis. *Protein science*, **20**, 1298–1345.

[4] Sriram, S. M., Kim, B. Y., and Kwon, Y. T. (2011) The N-end rule pathway: emerging functions and molecular principles of substrate recognition. *Nature reviews molecular cell biology*, **12**, 735–747.

[5] Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R., and Vandekerck hove, J. (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nature biotechnology*, **21**, 566–569.

[6] Xu, G., Shin, S. B. Y., and Jaffrey, S. R. (2009) Global profiling of protease cleavage sites by chemoselective labeling of protein N-termini. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 19310–19315.

[7] Nakazawa, T., Yamaguchi, M., Okamura, T.-a., Ando, E., Nishimura, O., and Tsunasawa, S. (2008) Terminal proteomics: N- and C-terminal analyses for high-fidelity identification of proteins using MS. *Proteomics*, **8**, 673–685.

[8] Xu, G., Shin, S. B. Y., and Jaffrey, S. R. (2011) Chemoenzymatic labeling of protein C-termini for positive selection of C-terminal peptides. *ACS Chemical Biology*, **6**, 1015–1020.

[9] Sanger, F. (1949) The terminal peptides of insulin. *Biochemical journal*, **45**, 563.

[10] Sanger, F. and Tuppy, H. (1951) The amino-acid sequence in the phenylalanyl chain of insulin. The identification of lower peptides from partial hydrolysates. *Biochemical journal*, **49**, 463.

[11] Sanger, F. and Thompson, E. (1953) The amino-acid sequence in the glycyl chain of insulin. The identification of lower peptides from partial hydrolysates. *Biochemical journal*, **53**, 353.

[12] Edman, P. (1950) Method for determination of the amino acid sequence in peptides. *Acta chemica Scandinavica*, **4**, 7.

[13] Kruft, V., Kapp, U., and Wittmannliebold, B. (1991) On-sequencer pyridylethylation of cysteine residues after protection of amino-groups by reaction with phenylisothiocyanate. *Analytical biochemistry*, **193**, 306–309.

[14] Iwamatsu, A. (1992) S-carboxymethylation of proteins transferred onto polyvinylidene difluoride membranes followed by *in situ* protease digestion and amino-acid microsequencing. *Electrophoresis*, **13**, 142–147.

[15] Ploug, M., Stoffer, B., and Jensen, A. L. (1992) Insitu alkylation of cysteine residues in a hydrophobic membrane-protein immobilized on polyvinylidene difluoride membranes by electroblotting prior to microsequence and amino-acid-analysis. *Electrophoresis*, **13**, 148–153.

[16] Jue, R. A. and Hale, J. E. (1993) Identification of cysteine residues alkylated with 3-bromopropylamine by protein-sequence analysis. *Analytical biochemistry*, **210**, 39–44.

[17] Krishna, R. G., Chin, C. C., and Wold, F. (1991) N-terminal sequence analysis of Nα-acetylated proteins after unblocking with N-acylaminoacyl-peptide hydrolase. *Analytical biochemistry*, **199**, 45–50.

[18] Hirano, H., Komatsu, S., Takakura, H., Sakiyama, F., and Tsunasawa, S. (1992) Deblocking and subsequent microsequence analysis of N-α-blocked proteins electroblotted onto pvdf membrane. *Journal of biochemistry*, **111**, 754–757.

[19] Bailey, J. M. (1995) Chemical methods of protein-sequence analysis. *Journal of chromatography A*, **705**, 47–65.

[20] Johnson, T. B. and Nicolet, B. H. (1911) Hydantoins: the synthesis of 2-thiohydantoin. *Journal of the American chemical society*, **33**, 1973–1978.

[21] Schlack, P. and Kampf, W. (1926) Über eine neue Methode zur Ermittlung der Konstitution von Peptiden. *Hoppe-Seyler´ s Zeitschrift für physiologische Chemie*, **154**, 125–172.

[22] Inglis, A. S. (1991) Chemical procedures for C-terminal sequencing of peptides and proteins. *Analytical biochemistry*, **195**, 183–196.

[23] Jones, J. et al. (1991) *Chemical synthesis of peptides*. Oxford Univ Press.

[24] Bailey, J. M., Shenoy, N. R., Ronk, M., and Shively, J. E. (1992) Automated carboxy-terminal sequence-analysis of peptides. *Protein science*, **1**, 68–80.

[25] Smith, B. J. (1997) *Protein sequencing protocols*, vol. 64. Springer.

[26] Nakazawa, T., Yamaguchi, M., Nishida, K., Kuyama, H., Obama, T., Ando, E., Okamura, T., Ueyama, N., Tanaka, K., and Norioka, S. (2004) Enhanced responses in matrix-assisted laser desorption/ionization mass spectrometry of peptides derivatized with arginine via a C-terminal oxazolone. *Rapid communications in mass spectrometry*, **18**, 799–807.

[27] Takamoto, K., Kamo, M., Kubota, K., Satake, K., and Tsugita, A. (1995) Carboxy-terminal degradation of peptides using perfluoroacyl anhydrides. *European journal of biochemistry*, **228**, 362–372.

[28] Stark, G. R. (1968) Sequential degradation of peptides from their carboxyl termini with ammonium thiocyanate and acetic anhydride. *Biochemistry*, **7**, 1796–1807.

[29] Yamaguchi, M., et al. (2006) Enhancement of MALDI-MS spectra of C-terminal peptides by the modification of proteins via an active ester generated *in situ* from an oxazolone. *Analytical chemistry*, **78**, 7861–7869.

[30] Bailey, J. M. and Shively, J. E. (1994) A chemical method for the C-terminal sequence analysis of proteins. *Methods*, **6**, 334 – 350.

[31] Miller, C. G. and Bailey, J. M. (1996) Automated C-terminal protein sequence analysis using the HP G1009A C-terminal protein sequencing system. *Hewlett-Packard journal*, **47**, 73–82.

[32] Kenner, G., Khorana, H., and Stedman, R. (1953) Selective removal of the C-terminal residue as a thiohydantoin. the use of diphenyl phosphorisothiocyanatidate. *Journal of the chemical society*, pp. 673–678.

[33] Bailey, J. M., Nikfarjam, F., Shenoy, N. R., and Shively, J. E. (1992) Automated carboxy-terminal sequence-analysis of peptides and proteins using diphenyl phosphoroisothiocyanatidate. *Protein science*, **1**, 1622–1633.

[34] Bailey, J. M., Tu, O., Issai, G., Ha, A., and Shively, J. E. (1995) Automated carboxy-terminal sequence-analysis of polypeptides containing C-terminal proline. *Analytical biochemistry*, **224**, 588–596.

[35] Miller, C. G., Hawke, D. H., Tso, J., and Early, S. (1995) Automated C-terminal protein sequence analysis using the Hewlett-Packard G1009A C-terminal protein sequencing system. *Hewlett-Packard journal*, **6**, 219–227.

[36] Boyd, V. L., Bozzini, M., Zon, G., Noble, R. L., and Mattaliano, R. J. (1992) Sequencing of peptides and proteins from the carboxy terminus. *Analytical biochemistry*, **206**, 344–352.

[37] Brune, D. C. (1992) Alkylation of cysteine with acrylamide for protein-sequence analysis. *Analytical biochemistry*, **207**, 285–290.

[38] Bozzini, M., Zhao, J. D., Yuan, P. M., Ciolek, D., Pan, Y. C., Horton, J., Marshak, D. R., and Boyd, V. L. (1995) Applications using an alkylation method for carboxy-terminal protein sequencing. *Techniques in protein chemistry*, **6**, 229–237.

[39] Samyn, B., Hardeman, K., Van der Eycken, J., and Van Beeumen, J. (2000) Applicability of the alkylation chemistry for chemical C-terminal protein sequence analysis. *Analytical chemistry*, **72**, 1389–1399.

[40] Boyd, V. L., Bozzini, M., Guga, P. J., Defranco, R. J., Yuan, P. M., Loudon, G. M., and Nguyen, D. (1995) Activation of the carboxy-terminus of a peptide for carboxy-terminal sequencing. *Journal of organic chemistry*, **60**, 2581–2587.

[41] Hardeman, K., Samyn, B., Beeumen, J. V., and Eycken, J. V. D. (1998) An improved chemical approach toward the C-terminal sequence analysis of proteins containing all natural amino acids. *Protein science*, **7**, 1593–1602.

[42] Aebersold, R., Bures, E. J., Namchuk, M., Goghari, M. H., Shushan, B., and Covey, T. C. (1992) Design, synthesis, and characterization of a protein sequencing reagent yielding amino-acid derivatives with enhanced detectability by mass-spectrometry. *Protein science*, **1**, 494–503.

[43] Chait, B. T., Wang, R., Beavis, R. C., and Kent, S. B. H. (1993) Protein ladder sequencing. *Science*, **262**, 89–92.

[44] Tsugita, A., Takamoto, K., K., Kamo, M., M., and Iwadate, H. (1992) C-terminal sequencing of protein. *European journal of biochemistry*, **206**, 691–696.

[45] Miyazaki, K. and Tsugita, A. (2004) C-terminal sequencing method for peptides and proteins by the reaction with a vapor of perfluoric acid in acetic anhydride. *Proteomics*, **4**, 11–19.

[46] Miyazaki, K. and Tsugita, A. (2006) C-terminal sequencing method for proteins in polyacrylamide gel by the reaction of acetic anhydride. *Proteomics*, **6**, 2026–2033.

[47] Zhong, H., Zhang, Y., Wen, Z., and Li, L. (2004) Protein sequencing by mass analysis of polypeptide ladders after controlled protein hydrolysis. *Nature biotechnology*, **22**, 1291–1296.

[48] Chen, L., Wang, N., Sun, D., and Li, L. (2014) Microwave-assisted acid hydrolysis of proteins combined with peptide fractionation and mass spectrometry analysis for characterizing protein terminal sequences. *Journal of proteomics*, **100**, 68–78.

[49] Schar, M., Bornsen, K. O., and Gassmann, E. (1991) Fast protein-sequence determination with matrix-assisted laser desorption and ionization mass-spectrometry. *Rapid communications in mass spectrometry*, **5**, 319–326.

[50] Patterson, D. H., Tarr, G. E., Regnier, F. E., and Martin, S. A. (1995) C-terminal ladder sequencing via matrix-assisted laser-desorption mass-spectrometry coupled with carboxypeptidase-Y time-dependent and concentration-dependent digestions. *Analytical chemistry*, **67**, 3971–3978.

[51] Samyn, B., Sergeant, K., Castanheira, P., Faro, C., and Van Beeumen, J. (2005) A new method for C-terminal sequence analysis in the proteomic era. *Nature methods*, **2**, 193–200.

[52] Samyn, B., Sergeant, K., and Beeumen, J. V. (2006) A method for C-terminal sequence analysis in the proteomic era (proteins cleaved with cyanogen bromide). *Nature protocols*, **1**, 317–322.

[53] Breddam, K. (1986) Serine carboxypeptidases. A review. *Carlsberg research communications*, **51**, 83–128.

[54] Selo, I., Negroni, L., Creminon, C., Grassi, J., and Wal, J. M. (1996) Preferential labeling of α-amino N-terminal groups in peptides by biotin, application to the detection of specific anti-peptide antibodies by enzyme immunoassays. *Journal of immunological methods*, **199**, 127–138.

[55] Kimmel, J. R. (1967) Guanidination of proteins. *Methods in enzymology*, **11**, 584–589.

[56] Beardsley, R. L., Karty, J. A., and Reilly, J. P. (2000) Enhancing the intensities of lysine-terminated tryptic peptide ions in matrix-assisted laser desorption/ionization mass spectrometry. *Rapid communications in mass spectrometry*, **14**, 2147–2153.

[57] Beardsley, R. L. and Reilly, J. P. (2002) Optimization of guanidination procedures for MALDI mass mapping. *Analytical chemistry*, **74**, 1884–1890.

[58] Keough, T., Lacey, M., and Youngquist, R. (2000) Derivatization procedures to facilitate *de novo* sequencing of lysine-terminated tryptic peptides using postsource decay matrix-assisted laser desorption/ionization mass spectrometry. *Rapid communications in mass spectrometry*, **14**, 2348–2356.

[59] Sergeant, K., Samyn, B., Debyser, G., and Van Beeumen, J. (2005) *De novo* sequence analysis of N-terminal sulfonated peptides after in-gel guanidination. *Proteomics*, **5**, 2369–2380.

[60] Dixon, H. (1964) Transamination of peptides. *Biochemical journal*, **92**, 661.

[61] Dixon, H. B. and Fields, R. (1972) Specific modification of NH_2-terminal residues by transamination. *Methods in enzymology*, **25**, 409–419.

[62] Wachter, E., Machleidt, W., Hofner, H., and Otto, J. (1973) Aminopropyl glass and its p-phenylene diisothiocyanate derivative, a new support in solid-phase Edman degradation of peptides and proteins. *FEBS letters*, **35**, 97–102.

[63] Gerstein, M. (1998) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Folding and design*, **3**, 497–512.

[64] Kim, J.-S., Shin, M., Song, J.-S., An, S., and Kim, H.-J. (2011) C-terminal *de novo* sequencing of peptides using oxazolone-based derivatization with bromine signature. *Analytical biochemistry*, **419**, 211–216.

[65] http://www.piercenet.com/product/nhs-sulfo-nhs.

[66] Kim, J.-S., Song, J.-S., Kim, Y., Park, S. B., and Kim, H.-J. (2012) *De novo* analysis of protein N-terminal sequence utilizing MALDI signal enhancing derivatization with Br signature. *Analytical and bioanalytical chemistry*, **402**, 1911–1919.

[67] Shin, M. and Kim, H.-J. (2011) Peptide C-terminal sequence analysis by MALDI-TOF MS utilizing EDC coupling with Br signature. *Bulletin of the Korean chemical society*, **32**, 1183.

[68] Kosaka, T., Takazawa, T., and Nakamura, T. (2000) Identification and C-terminal characterization of proteins from two-dimensional polyacrylamide gels by a combination of isotopic labeling and nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry. *Analytical chemistry*, **72**, 1179–1185.

[69] Julka, S., Dielman, D., and Young, S. A. (2008) Detection of C-terminal peptide of proteins using isotope coding strategies. *Journal of chromatography B*, **874**, 101–110.

[70] Stewart, I., Thomson, T., and Figeys, D. (2001) O-18 labeling: a tool for proteomics. *Rapid communications in mass spectrometry*, **15**, 2456–2465.

[71] Murphy, C. M. and Fenselau, C. (1995) Recognition of the carboxy-terminal peptide in cyanogen-bromide digests of proteins. *Analytical chemistry*, **67**, 1644–1645.

[72] Nakajima, C., Kuyama, H., Nakazawa, T., and Nishimura, O. (2012) C-terminal sequencing of protein by MALDI mass spectrometry through the specific derivatization of the α-carboxyl group with 3-aminopropyltris-(2, 4, 6-trimethoxyphenyl) phosphonium bromide. *Analytical and bioanalytical chemistry*, **404**, 125–132.

[73] McDonald, L., Robertson, D. H. L., Hurst, J. L., and Beynon, R. J. (2005) Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nature methods*, **2**, 955–957.

[74] Yamaguchi, M., et al. (2007) Specific isolation of N-terminal fragments from proteins and their high-fidelity *de novo* sequencing. *Rapid communications in mass spectrometry*, **21**, 3329–3336.

[75] Timmer, J., et al. (2007) Profiling constitutive proteolytic events *in vivo*. *Biochemical journal*, **407**, 41–48.

[76] Yamaguchi, M., et al. (2008) Selective isolation of N-terminal peptides from proteins and their *de novo* sequencing by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry without regard to unblocking or blocking of N-terminal amino acids. *Rapid communications in mass spectrometry*, **22**, 3313–3319.

[77] Shen, P.-T., Hsu, J.-L., and Chen, S.-H. (2007) Dimethyl isotope-coded affinity selection for the analysis of free and blocked N-termini of proteins using LC-MS/MS. *Analytical chemistry*, **79**, 9520–9530.

[78] Bailon, P., Ehrlich, G. K., Fung, W.-J., and Berthold, W. (2000) Affinity chromatography, methods and protocols. *Methods in molecular biology*.

[79] Kuyama, H., Shima, K., Sonomura, K., Yamaguchi, M., Ando, E., Nishimura, O., and Tsunasawa, S. (2008) A simple and highly successful C-terminal sequence analysis of proteins by mass spectrometry. *Proteomics*, **8**, 1539–1550.

[80] Sonomura, K., Kuyama, H., Matsuo, E.-i., Tsunasawa, S., and Nishimura, O. (2009) The specific isolation of C-terminal peptides of proteins through a transamination reaction and its advantage for introducing functional groups into the peptide. *Rapid communications in mass spectrometry*, **23**, 611–618.

[81] Kuyama, H., Nakajima, C., and Tanaka, K. (2012) Enriching C-terminal peptide from endopeptidase ArgC digest for protein C-terminal analysis. *Bioorganic & medicinal chemistry letters*, **22**, 7163–7168.

[82] Nika, H., Nieves, E., Hawke, D. H., and Angeletti, R. H. (2013) C-terminal protein characterization by mass spectrometry using combined micro scale liquid and solid-phase derivatization. *Journal of biomolecular techniques: JBT*, **24**, 17.

[83] Nika, H., Hawke, D. H., and Angeletti, R. H. (2014) C-terminal protein characterization by mass spectrometry: isolation of C-terminal fragments from cyanogen bromide-cleaved protein. *Journal of biomolecular techniques: JBT*, **25**, 1.

[84] Sechi, S. and Chait, B. (2000) A method to define the carboxyl terminal of proteins. *Analytical chemistry*, **72**, 3374–3378.

[85] Mahrus, S., Trinidad, J. C., Barkan, D. T., Sali, A., Burlingame, A. L., and Wells, J. A. (2008) Global sequencing of proteolytic cleavage sites in apoptosis by specific labeling of protein N-termini. *Cell*, **134**, 866–876.

[86] Dormeyer, W., Mohammed, S., van Breukelen, B., Krijgsveld, J., and Heck, A. J. R. (2007) Targeted analysis of protein termini. *Journal of proteome research*, **6**, 4634–4645.

[87] Armengaud, J. (2009) A perfect genome annotation is within reach with the proteomics and genomics alliance. *Current opinion in microbiology*, **12**, 292–300.

[88] Mohammed, S. and Heck, A. J. (2011) Strong cation exchange (SCX) based analytical methods for the targeted analysis of protein post-translational modifications. *Current opinion in biotechnology*, **22**, 9–16.

[89] Gauci, S., Helbig, A. O., Slijper, M., Krijgsveld, J., Heck, A. J., and Mohammed, S. (2009) Lys-N and trypsin cover complementary parts of the phosphoproteome in a refined SCX-based approach. *Analytical chemistry*, **81**, 4493–4501.

[90] Gorman, J. J. and Shiell, B. J. (1993) Isolation of carboxyl-termini and blocked amino-termini of viral proteins by high-performance cation-exchange chromatography. *Journal of chromatography A*, **646**, 193–205.

[91] Brown, J. and Hartley, B. (1966) Location of disulphide bridges by diagonal paper electrophoresis. *Biochemical journal*, **101**, 214–228.

[92] Cruickshank, W. H., Malchy, B. L., and Kaplan, H. (1974) Diagonal chromatography for the selective purification of tyrosyl peptides. *Canadian journal of biochemistry*, **52**, 1013–1017.

[93] Duggleby, R. G. and Kaplan, H. (1975) A general method for the determination of the carboxyl-terminal sequence of proteins. *Analytical biochemistry*, **65**, 346–354.

[94] Bietlot, H. P., Carey, P. R., Pozsgay, M., and Kaplan, H. (1989) Isolation of carboxyl-terminal peptides from proteins by diagonal electrophoresis: Application to the entomocidal toxin from *Bacillus thuringiensis*. *Analytical biochemistry*, **181**, 212–215.

[95] Gevaert, K., Van Damme, J., Goethals, M., Thomas, G. R., Hoorelbeke, B., Demol, H., Martens, L., Puype, M., Staes, A., and Vandekerckhove, J. (2002) Chromatographic isolation of methionine-containing peptides for gel-free proteome analysis identification of more than 800 *Escherichia coli* proteins. *Molecular & cellular proteomics*, **1**, 896–903.

[96] Gygi, S., Rist, B., Gerber, S., Turecek, F., Gelb, M., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature biotechnology*, **17**, 994–999.

[97] Gevaert, K., Ghesquière, B., Staes, A., Martens, L., Van Damme, J., Thomas, G. R., and Vandekerckhove, J. (2004) Reversible labeling of cysteine-containing peptides allows their specific chromatographic isolation for non-gel proteome studies. *Proteomics*, **4**, 897–908.

[98] Gevaert, K., Staes, A., Van Damme, J., De Groot, S., Hugelier, K., Demol, H., Martens, L., Goethals, M., and Vandekerckhove, J. (2005) Global phosphoproteome analysis on human HepG2 hepatocytes using reversed-phase diagonal LC. *Proteomics*, **5**, 3589–3599.

[99] Ghesquière, B., Van Damme, J., Martens, L., Vandekerckhove, J., and Gevaert, K. (2006) Proteome-wide characterization of N-glycosylation events by diagonal chromatography. *Journal of proteome research*, **5**, 2438–2447.

[100] Ghesquière, B., Buyl, L., Demol, H., Van Damme, J., Staes, A., Timmerman, E., Vandekerckhove, J., and Gevaert, K. (2007) A new approach for mapping sialylated N-glycosites in serum proteomes. *Journal of proteome research*, **6**, 4304–4312.

[101] Hanoulle, X., Van Damme, J., Staes, A., Martens, L., Goethals, M., Vandekerckhove, J., and Gevaert, K. (2006) A new functional, chemical proteomics technology to identify purine nucleotide binding sites in complex proteomes. *Journal of proteome research*, **5**, 3438–3445.

[102] Stes, E., Laga, M., Walton, A., Samyn, N., Timmerman, E., De Smet, I., Goormachtig, S., and Gevaert, K. (2014) A COFRADIC protocol to study protein ubiquitination. *Journal of proteome research*.

[103] Staes, A., Van Damme, P., Helsens, K., Demol, H., Vandekerckhove, J., and Gevaert, K. (2008) Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC). *Proteomics*, **8**, 1362–1370.

[104] Van Damme, P., Staes, A., Bronsoms, S., Helsens, K., Colaert, N., Timmerman, E., Aviles, F. X., Vandekerckhove, J., and Gevaert, K. (2010) Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nature methods*, **7**, 512–515.

[105] Sandra, K., Verleysen, K., Labeur, C., Vanneste, L., D'Hondt, F., Thomas, G., Kas, K., Gevaert, K., Vandekerckhove, J., and Sandra, P. (2007) Combination of COFRADIC and high temperature - extended column length conventional liquid chromatography: A very efficient way to tackle complex protein samples, such as serum. *Journal of separation science*, **30**, 658–668.

[106] Staes, A., Impens, F., Van Damme, P., Ruttens, B., Goethals, M., Demol, H., Timmerman, E., Vandekerckhove, J., and Gevaert, K. (2011) Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nature protocols*, **6**, 1130–1141.

[107] Schilling, O., Barre, O., Huesgen, P. F., and Overall, C. M. (2010) Proteome-wide analysis of protein carboxy termini: C terminomics. *Nature methods*, **7**, 508–U33.

[108] Kleifeld, O., Doucet, A., Keller, U. A. D., Prudova, A., Schilling, O., Kainthan, R. K., Starr, A. E., Foster, L. J., Kizhakkedathu, J. N., and Overall, C. M. (2010) Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nature biotechnology*, **28**, 281–U144.

[109] Helbig, A. O., Gauci, S., Raijmakers, R., van Breukelen, B., Slijper, M., Mohammed, S., and Heck, A. J. R. (2010) Profiling of N-acetylated protein termini provides in-depth insights into the N-terminal nature of the proteome. *Molecular & cellular proteomics*, **9**, 928–939.

[110] Lange, P. F., Huesgen, P. F., Nguyen, K., and Overall, C. M. (2014) Annotating N-termini for the human proteome project: N termini and N-α-acetylation status differentiate stable cleaved protein species from degradation remnants in the human erythrocyte proteome. *Journal of proteome research*, **13**, 2028–2044.

[111] Fahlman, R. P., Chen, W., and Overall, C. M. (2014) Absolute proteomic quantification of the activity state of proteases and proteolytic cleavages using proteolytic signature peptides and isobaric tags. *Journal of proteomics*, **100**, 79–91.

Rationale and aims

When this project was started in 2007, the ladder sequencing approach introduced in section 2.3.1, using carboxypeptidase (Cpase) digestion on CNBr generated peptides, was the only C-terminal sequencing method that had successfully been applied at a proteomic scale and that was used to study proteolytic processes in biological samples [1, 2]. Unfortunatly, the technique has some important limitations. The main goals of this research were to overcome those limitations and to automate the improved method, so that it can be applied to complex biological samples in a high-throughput setup.

The first aim was to develop a chemical method to identify the C-terminal peptide in a mixture of CNBr generated peptides, as an alternative to the use of CPase. CPase cleaves off different amino acids at different rates, while certain amino acids are not cleaved off at all. Therefore, to obtain optimal results, the digestion protocol must be specifically optimized for every peptide. In our chemical selection method, CNBr-generated peptides are incubated in a slightly basic buffer. All homoserine lactone residues, present in internal and N-terminal peptides, are partially opened and form homoserine, and are displayed as a doublet in mass spectra. This allows to discriminate the C-terminal peptide that can then be identified by MS/MS. This allowed us to set up the first fully automated C-terminal sequencing platform that has been tested in a traditional proteomic setting on *Shewanella oneidensis* MR-1. The development, automation and application of the chemical selection technique will be discussed in Chapter 3.

Since C-terminal fragments are identified using a MALDI TOF/TOF setup, the C-terminal fragments have to be smaller than 5 kDa in order to be detected with sufficient resolution. Since methionine only accounts for 2.59% of the amino acids in *Shewanella oneidensis* MR-1, the peptides generated after CNBr cleavage are often relatively large. Generating smaller fragments results in a higher number of proteins for which the C-terminal peptide can be analyzed. The second goal of our study was to generate smaller C-terminal peptides using a new chemical method that still retains the capability to select the C-terminal peptide in the mixture. We present a new cleavage method that cleaves C-terminal of Met and Trp simultaneously by adding KI to the standard CNBr cleavage mixture. During the cleavage reaction all Trp residues are converted to Cγ-O-spirolactone tryptophan. We further showed that the chemical selection technique to discriminate the C-terminal peptide can also be applied to CNBr and KI generated

digest mixtures. The alternative cleavage reaction increased the theoretical coverage of the *Shewanella oneidensis* MR-1 proteome to 58%, almost 10% higher than can be achieved using any other standard cleavage methods. The generation of smaller C-terminal fragments by cleavage after Met and Trp using a CNBr and KI mixture are discussed in Chapter 4.

In the course of 2010 the groups of Gevaert and Overall reported two new proteome-wide LC-MS based C-terminal sequencing techniques [3, 4]. The techniques were applied in multiple degradomics and terminomics studies identifying a few hundred new C- and N-termini for each study. This increase in identification and the low ionization and fragmentation efficiency of C-terminal peptides by MALDI TOF/TOF forced us to develop a completely new setup, that can be analyzed on any (LC-)MS platform, to become competitive in high-throughput terminomics. In this new approach all carboxyl groups of the intact protein are derivatized using 1-(2-pyrimidyl)piperazine (PP), a method originally developed to improve the ionization efficiency in phosphopeptides. This carbodiimide-mediated coupling reaction is followed by a proteolytic digest step. After digestion, the free carboxyl groups of the internal and N-terminal peptides are bound to a polymer, leaving the C-terminal peptides in solution. The PP-group at the original C-terminus serves as a positive selection procedure to eliminate false positives, and should at the same time improve their ionization efficiency. The preliminary results obtained using the PP derivatization and enrichment are presented in Chapter 5.

References

[1] Samyn, B., Sergeant, K., Castanheira, P., Faro, C., and Van Beeumen, J. (2005) A new method for C-terminal sequence analysis in the proteomic era. *Nature methods*, **2**, 193–200.

[2] Samyn, B., Sergeant, K., and Beeumen, J. V. (2006) A method for C-terminal sequence analysis in the proteomic era (proteins cleaved with cyanogen bromide). *Nature protocols*, **1**, 317–322.

[3] Van Damme, P., Staes, A., Bronsoms, S., Helsens, K., Colaert, N., Timmerman, E., Aviles, F. X., Vandekerckhove, J., and Gevaert, K. (2010) Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nature methods*, **7**, 512–515.

[4] Schilling, O., Barre, O., Huesgen, P. F., and Overall, C. M. (2010) Proteome-wide analysis of protein carboxy termini: C terminomics. *Nature methods*, **7**, 508–U33.

Part II

Results

Chapter 3

Chemical selection of C-terminal peptide after CNBr digest

3.1 Introduction

In the carboxypeptidase (CPase) based ladder sequencing technique, already introduced in section 2.3.1, proteins are chemically cleaved using cyanogen bromide (CNBr). During CNBr cleavage, all Met-Xxx peptide bonds are cleaved and all methionine residues are converted to a homoserine-derivative, which is in equilibrium with its lactone form. During digestion with carboxypeptidases only the original C-terminal fragment (having a free carboxyl group) is accessible to enzymatic degradation and forms a ladder. We present here a novel chemical selection method where a slightly basic buffer treatment is used to partially open the homoserine lactone forming homoserine. This results in the formation of m/z doublets (Δm = 18 Da) for all internal peptides and allows to identify the C-terminal peptide, which appears as singlet in the mass spectrum. The sequence of the C-terminal peptide is determined by tandem MS using a MALDI TOF/TOF instrument.

3.1.1 CNBr cleavage of proteins

In 1962, Gross and Witkop first proposed the use of CNBr as cleavage agent for peptide bonds in proteins [1]. The selectivity of the reaction depends on the pH. In neutral and alkaline conditions, CNBr reacts with basic groups in proteins [2]. At acidic pH only Met and free Cys are attacked by CNBr, and cysteine is oxidized to cysteic acid [3]. At low pH the cleavage yield is nearly 100%, except if the residue C-terminal of methionine is serine or threonine [4]. The hydroxyl function of these amino acids can interfere with the cleavage reaction by acting as nucleophile and competes with water in the cleavage reaction (Figure 3.1). This side reaction can be avoided by performing the reaction in a solution containing a higher percentage

of water [5]. Incubation with 70% TFA is preferred, as formylation of reactive side chains is observed if 70% formic acid is used [6, 7]. Despite the harsh reaction conditions, in situ CNBr cleavage of proteins in-gel has been demonstrated [8, 9]. It is important to note that oxidation of methionine totally impairs the reaction.

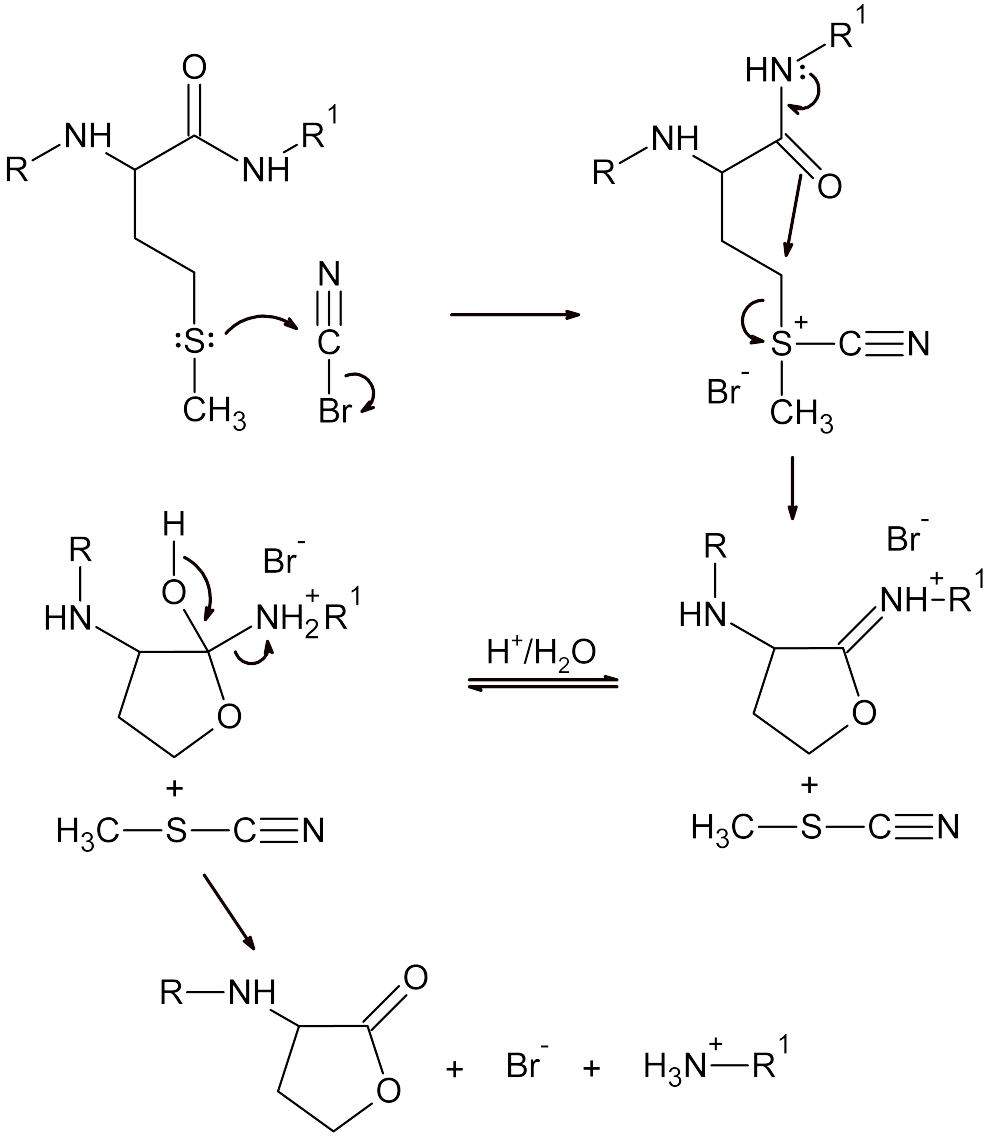


Figure 3.1: CNBr cleavage reaction. Incubation of proteins with CNBr results in cleavage of peptide bonds C-terminal to methionine. During the cleavage reaction methionine is converted to a homoserine lactone. The acid stable homoserine lactone is in equilibrium with its open form (homoserine). In the first step, the sulfur function of methionine displaces the polarized bromide anion of cyanogen bromide forming a bromide-stabilized S-cyanide methionine derivate. In the second step the nucleophilic attack of the peptide bond carbonyl function results in the formation of a five membered ring structure and the release of methylisothiocyanate. This iminolactone ring is resolved by nucleophilic attack from water, resulting in peptide bond cleavage and the formation of homoserine lactone [5].

The equilibrium between the homoserine lactone and its open form is pH sensitive. In acidic environment the lactone form is preferred over the open form [10]. Under CNBr cleavage reaction conditions (70% TFA) and CPase incubation buffer (10 mM ammonium acetate pH 5.4) the equilibrium is almost completely shifted to the lactone form.

The specific reactivity of the lactone function has been used in different approaches. The first report on homoserine lactone aminolysis with aryl and alkylamines dates back to 1964 [11]. In solid-phase Edman degradation the lactone ring is used to couple the CNBr-generated peptides to an insoluble amino resin [12]. More recently, the same chemistry was used to conjugate amino group containing affinity tags to homoserine lactone ending peptides, as a method to identify disulfide bridges and chemical cross-linking sites [13]. Likewise the addition of tris(hydroxymethyl)aminomethane (Tris) was used in order to improve the mass accuracy determination of high molecular mass CNBr generated peptides [14].

The homoserine lactone fragments have also been used to identify the C-terminal fragment in a CNBr-generated peptide mixture. Lactonization and amination with radioactive [^{14}C]ethylenediamine of CNBr-generated fragments allowed to identify the C-terminal peptide after peptide separation [15]. Reaction of acidic methanol with homoserine lactone causes a 32 Da mass shift for all internal peptides. The C-terminal peptide can be identified when the pre- and post-derivatization MS spectra are compared [16].

3.1.2 Carboxypeptidase methodology

Two carboxypeptidases, P and Y, are used to determine the C-terminal peptide in a CNBr digest. They are both serine exoproteases with a broad substrate specificity [17–19]. CPY is a vacuolar protease isolated and characterized from *Saccharomyces cerevisiae*, and CPP was isolated from *Penicillium janthinellum* [20]. A unique feature of these enzymes is their wide pH optimum and, in particular, their high activity at acidic pH: between 2.5 and 5.7 for CPP and between 5.5 and 7.5 for CPY [21]. This is in contrast to members of the trypsin or subtilisin families of serine endopeptidases, which are essentially inactive at pH <7 [22, 23].

Structural studies of CPY show that the catalytic mechanism resembles that of typical serine proteases, which hydrolyze peptide and ester substrates in a two-step reaction using a catalytic triad (Figure 3.2). A key feature of this mechanism is the ability of the catalytic histidine to abstract a proton from the catalytic serine prior to, or simultaneously with, the nucleophilic attack on the scissile bond. Several models explain how the catalytically essential histidine is maintained in its active deprotonated state through perturbation of its pK_a-value in the enzyme-substrate complex [24, 25]. One model involves the polarization of the scissile peptide

bond through H-bond formation with the α-carboxylate group. According to a second model an apolar environment can be formed in the catalytic site upon binding of the substrate [19].

CPY has six substrate binding sites S1' to S5, of which S1 is the most selective one and a binding site for the carboxylate group [26]. This explains why despite their broad specificity, preferences for both the C-terminal and the penultimate amino acid exist. Hydrophobic residues are preferred at the penultimate position; Phe>Leu>Ala>Ile>>>Lys for both carboxypeptidases [21]. CPP has a lower cleavage rate for Ser and Gly as C-terminal residue, while incubation with CPY alone results in a slower cleavage rate when Asp and Gly are the C-terminal amino acids. Incubation with a combination of CPP and CPY overcomes the specificity of the individual carboxypeptidases. Longer stretches of sequence can be obtained if a combination of these carboxypeptidases is used [27].

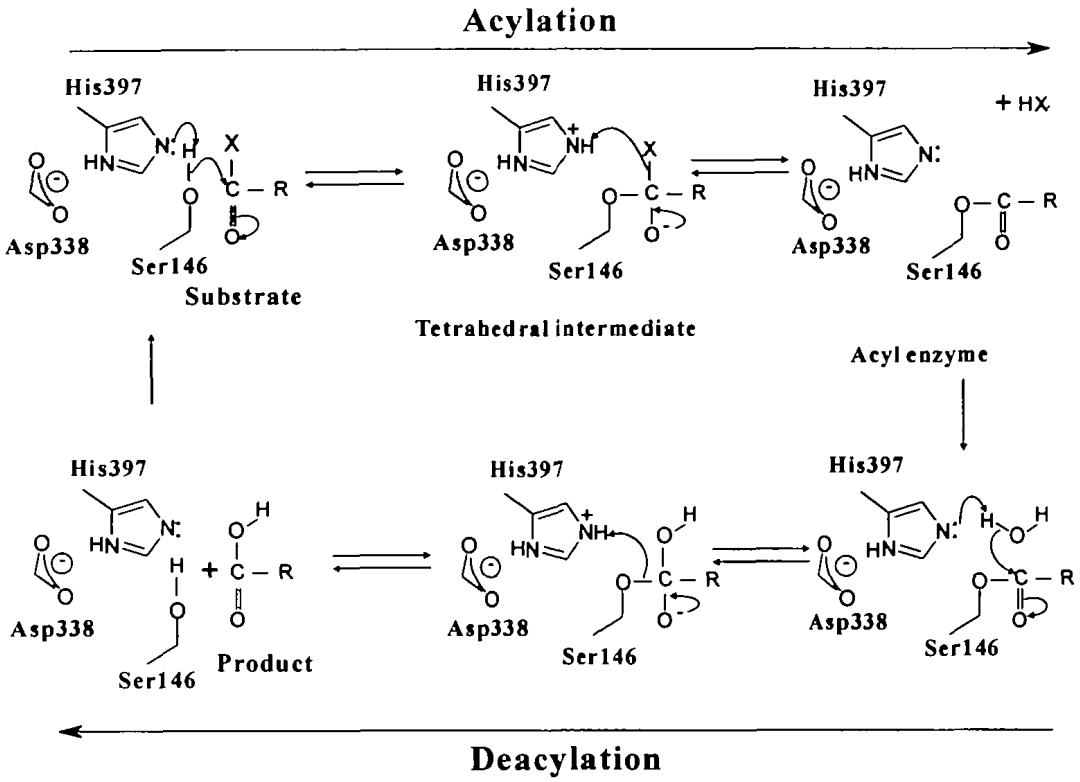


Figure 3.2: The catalytic mechanism of CPY (typical for serine proteases), which hydrolyzes peptide and ester substrates in a two-step reaction. In the first reaction, a tetrahedral intermediate is formed as a result of the nucleophilic attack of the essential serine hydroxyl on the carbonyl carbon atom of the substrate. The histidine residue assists in this step by accepting the proton from the serine hydroxyl and stabilizing the tetrahedral intermediate. This proton is then transferred to the newly generated amino half of the peptide bond in an acylation step. The amino half of the peptide is now free to dissociate. Hydrolysis of the acyl-enzyme intermediate then produces a product in a deacylation step [19].

Because CPases cleave off different amino acids at different rates, with certain amino acids not being cleaved off at all, the digestion protocol should be specifically optimized for every peptide. In order to obtain a full sequence ladder, analysis at multiple time points or at different concentrations of CPase need to be acquired, and MS spectra need to be combined. This makes the approach labor intensive. Since the peptide of interest is distributed over multiple experiments and present in multiple ladder forms simultaneously, the sensitivity of the technique is low.

Therefore, we designed a novel strategy to make the C-terminal sequencing method less dependent on CPase activity. We designed a chemical selection method where a slightly basic buffer is used that partially opens the homoserine lactone, forming homoserine. This results in the formation of m/z doublets (Δm = 18 Da) for all internal peptides and allows to identify the C-terminal peptide, which appears as singlet in the mass spectrum. The sequence of the C-terminal peptide is determined by tandem MS using a MALDI TOF/TOF mass spectrometer. The techniques can both be applied on purified proteins in solution and in gel (1D and 2D PAGE). Since all the homoserine lactone residues open under the same conditions, one protocol can be applied to all samples. This allows automation of the technique. Here we present results of proof-of-concept experiments and an automation of the procedure on a robotic Tecan platform.

In this study the entire cell lysate of *Shewanella oneidensis* MR-1 was used as a test case. *S. oneidensis* is a Gram-negative γ-proteobacterium isolated from freshwater lake sediment [28]. The organism is an aerotolerant anaerobe, able to reduce heavy metal ions, making it a potential agent for bioremediation. The *S. oneidensis* MR-1 genome has been sequenced in 2002 [29] and has been reannotated in 2003 [30]. Multiple proteomics studies have attemped to use LC-MS and LC-MS/MS data to improve the annotations of *S. oneidensis* [31–34] and to detect post-translational modifications such as proteolytical processing [35]. Recently a large mutant study was performed in order to improve the annotation of gene function of *S. oneidensis* proteins [36].

This chemical selection technique has recently been used to determine the C-terminal sequence of the recombinant protein α-1-antitrypsin (A1PI) after expression in a human cell line. α-1-antitrypsin is a plasma protein with the function to protect lung tissues from proteolytic destruction by enzymes from inflammatory cells. A1PI deficiency is an inherited disorder associated with pulmonary emphysema and a higher risk of chronic obstructive pulmonary disease [37].

3.2 A new chemical approach to differentiate carboxy terminal peptide fragments in cyanogen bromide digests of proteins.

P.P. Moerman[1], K. Sergeant[2], G. Debyser[1], B. Devreese[1], B. Samyn[1*]

[1] Laboratory for Protein Biochemistry and Biomolecular Engineering, Ghent University, Ghent, Belgium
[2] Department Environment and Agro-biotechnologies, Centre de Recherche Public, Gabriel Lippmann, Belvaux, Luxembourg

*Corresponding author: Bart.samyn@ugent.be

This chapter has been published in Journal of Proteomics (2010), 73(8), 1454-1460

3.2.1 Abstract

We present a novel approach to perform C-terminal sequence analysis by discerning the C-terminal peptide in a mass spectral analysis of a CNBr digest. During CNBr cleavage, all Met-Xxx peptide bonds are cleaved and the generated internal peptides all end with a homoserine lactone (hsl) derivative. The partial opening of the hsl derivatives, by using a slightly basic buffer solution, results in the formation of m/z doublets (Δm = 18 Da) for all internal peptides, and allows to identify the C-terminal peptide which appears as a singlet in the mass spectra. Using two model proteins we demonstrate that this approach can be applied to study proteins purified in gel or in solution. The chemical opening of the hsl derivative does not require any sample clean-up and therefore, the sensitivity of the C- terminal sequencing approach is increased significantly. Finally, the new protocol was applied to characterize the C-terminal sequence of two recombinant proteins. Tandem mass spectrometry by MALDI-TOF/TOF allowed to identify the sequence of the C-terminal peptides. This novel approach will allow to perform a proteome-wide study of C-terminal proteolytic processing events in a high-throughput fashion.

3.2.2 Introduction

Proteomics is one of the fastest growing areas of biological research. Its objective has moved beyond simple cataloguing to the study of functional and regulatory aspects of proteins. The monitoring of protein expression profiles remains a very challenging task because of the wide dynamic range of expressed proteins and the variability of gene products due to the presence of splicing variants, N- and C-terminal processing, and co- and post-translational modifications (PTMs). Truncations of the nascent polypeptide chain at the N- or C-terminus are by far the most common types of PTMs found in proteins. Although several approaches have been developed that allow proteome-wide PTM analysis of phosphorylation and glycosylation events, relatively little attention has been paid to the development of approaches for the systematic analysis of proteolytic processing events [38]. Recently, a number of methods have been developed to characterize N-terminal proteolytic processing, such as e.g. the COFRADIC [39] and the positional proteomics approach [40]. In contrast, considerably less attention has been spent on the study of C-terminal processing [41].

Recently, we developed a new technique that enables the more systematic identification of C-termini from intact proteins. This MS-based, enzymatic, ladder sequencing approach can be applied on the unseparated peptide mixture generated by cleavage of the protein with cyanogen bromide (CNBr), either in solution or in gel. Under acidic conditions, CNBr cleaves after every Met in the sequence, generating internal peptides with a C-terminal homoserine lactone (hsl) derivative. During incubation with carboxypeptidases (CPase) only the original

C-terminal fragment (having a free carboxyl group) is accessible to enzymatic degradation and forms a ladder. Ladder readout is typically performed using MALDI-TOF MS as this technique produces predominant ladders of singly charged ions. Application of this method on complex mixtures has verified its vital role for the determination of C-termini of proteins at a proteomic scale [42]. A positive identification of the C-terminus will first depend on the length and ionization capacity of the generated CNBr fragments. In our approach, MS analysis was performed using MALDI ionization which generates predominantly singly charged ions. The use of the α-cyano-4-hydroxycinnamic acid matrix produces interference in the low Mw mass range and therefore presents a challenge for the analysis of peptides with a Mw below 0.7-1.3 kDa. In our experience, the upper mass limit for the analysis of ladder sequences in RE-MALDI-MS, providing enough resolution and accuracy to identify amino acids by 0.1 Da weight difference, is restricted to C-terminal fragments with a Mw of $\pm$4 kDa.

In its current form the use of CPase limits the sensitivity and the specificity of this approach [43]. The use of exopeptidases has regained interest with the introduction of improved methods for MS readout. For C-terminal sequence analysis, carboxypeptidases Y and P (CPY & CPP) are chosen most often because of their broad amino acid specificity [44]. However, the rate at which amino acids are cleaved by CPase depends to a great extent on the amino acid sequence of the substrate. Additionally, the cleavage rate depends on reaction conditions such as pH, ion strength and substrate concentration. During digestion with CPase we observed, according to the known specificity, a slow cleavage of C-terminal Gly. Unexpectedly, the presence of Phe, Thr, or Lys also slowed down or inhibited ladder generation (unpublished results). Finally, it should be noted that most exopeptidases have a K_m-value in the range of 5-50 μM, which means that they are operating at 50% maximum velocity when a protein concentration of 5 pmol/μl is used [45]. At lower concentrations (sub picomole), the proteolytic activity toward the substrate will be minimal.

Therefore, the main goal of the present study was to eliminate the use of CPase and develop a novel chemical approach that allows to differentiate between internal CNBr fragments and the C-terminal peptide fragment. Incubation of the CNBr mixtures in a slightly basic buffer results in a partial opening of the homoserine lactone derivatives to the corresponding homoserine derivative (Δm = + 18 Da). Therefore, all internal peptides appear as doublets in the MS spectrum, whereas the C-terminal fragment appears as the only singlet and can be selected for MS/MS analysis. The feasibility of this approach was evaluated using two test proteins and demonstrated with two recombinant proteins purified in solution or by SDS-PAGE. We further demonstrate that, in contrast to the CPase method, the sensitivity of this approach is in the lower fmol range (150 fmol).

Previously, we have demonstrated that C-terminal sequence information is useful to study C-terminal processing events, but that its positional bias can also be used to improve the correctness of protein identification by mining protein databases [42]. As the N- and C-terminal sequences are highly specific characteristic features, this information can also be used to mine annotated as well as un-annotated genome sequences to identify ORFs. It has been demonstrated earlier that proteomic data information is an excellent tool to mine genome sequences, eliminating problems associated with standard techniques for gene prediction, especially with eukaryotic data [46]. So far, only high-throughput tandem MS based proteomics has been used to improve the quality of genome annotations [47]. Recently, the group of Heck has demonstrated that there are a significant number of unknown protein N- and C-termini in the human genome, suggesting the existence of a high degree of novel transcription, independent of annotated gene boundaries and/or specific protein processing [48]. The C-terminal sequence information will allow integrating proteomic MS data with current genomic annotations and thus improve the quality of gene prediction.

3.2.3 Materials and methods

Materials

Yeast alcohol dehydrogenase (gel filtration purity) and horse heart cytochrome c (purity 99%) were purchased from Sigma (Bornem, Belgium). Elongation factor Tu (EF-Tu or TufA) (*Shewanella oneidensis* MR-1) and glycerophosphodiester phosphodiesterase (*Haemophilus influenzae*) were expressed and purified in house. Stock solutions (0.5 nmol/μl water) were prepared for all proteins and further diluted prior to use. HPLC- grade acetonitrile (ACN) was obtained from BioSolve (Valkenswaard, The Netherlands). Trifluoroacetic acid (TFA) (purity >99.9%) was from Beckman Instruments (Palo Alto, CA, U.S.A.). Ammonium hydrogen carbonate (NH_4HCO_3) (purity >99.5%), ammonium citrate (purity 98%), cyanogen bromide (CNBr) (purity 97%) and α-cyano-4-hydroxycinnamic acid were purchased from Sigma. 12% Tris-HCl precast SDS-gels were purchased from Bio-Rad (Nazareth, Belgium). C-18 ZipTips were from Millipore (Billerica, MA, USA). Water was purified using a MilliQ water filtration system (Millipore). SDS- PAGE gels were performed as described previously [42, 43].

CNBr Cleavage

CNBr cleavage of gel/gel-free separated proteins was basically performed as described before [42, 43]. Briefly, after visualization and destaining, the bands or spots containing the protein were excised from the gels. Before cleavage with CNBr, the gel pieces were washed twice with 150 μl 50% ACN/MQ, dehydrated with 40 μl ACN and reswollen in 5 μl MQ. CNBr cleavage was started by adding 5 μl 5.0 M CNBr/ACN (Sigma, Bornem, Belgium) and 15 μl TFA (Applied Biosystems, Foster City, CA). CNBr and TFA are highly toxic and corrosive products

which must, at any time, be manipulated under a fumehood by skilled personnel wearing protective clothing! After incubation overnight (4 °C) the supernatant was collected and the peptides were extracted twice with 50 μl 70% ACN/0.1% TFA for 30 min at 37 °C. All fractions were pooled and dried in a SpeedVac. Care must be taken during sample manipulation to avoid oxidation. Artefactual oxidation of methionine to sulfoxide means that CNBr is inhibited from reacting with the sulfur of the methionine residue, which may explain the abscence of certain CNBr fragments.

Chemical opening

Before chemical opening, the dried samples were desalted using C-18 micro purification tips (ZipTip). The ZipTip protocol was performed as described by the manufacturer. A 50% ACN/ 0.1% TFA solution was used as activation and elution solvent. A 0.1% TFA solution in MQ was used as equilibration and washing solvent. After elution the samples were dried in a SpeedVac instrument. The partial opening of the homoserine lactone to homoserine was performed by redissolving the sample in 10 μl 12.5 mM NH_4HCO_3 (pH 8.0), and incubation at 37 °C for 30 min. Care was taken during the redissolving steps to thoroughly vortex the samples for 30 seconds.

Mass spectrometry

All mass spectrometric analysis were performed on an Applied Biosystems 4800+ Proteomics Analyzer with TOF/ TOF optics (Applied Biosystems, Foster City, CA). This mass spectrometer uses a 200-Hz frequency tripled Nd:YAG laser operating at a wavelength of 355 nm. For MS/MS, ions generated by the MALDI process are accelerated at 8 kV through a grid set at 7.3 kV into a short, linear, field-free drift region. In this region, the ions pass through a timed-ion selector device that is able to select one peptide, from a mixture of peptides, for subsequent fragmentation in the collision cell. After a peptide at a given m/z is selected by the timed-ion-selector, it passes through a retarding lens, where the ions are decelerated and then pass into the collision cell, which is operated at 7 kV. The collision energy is defined by the potential difference between the source and the collision cell (1 kV). After passing through the collision cell, the ions (both intact peptide ion and fragments) are accelerated in the second source region at 15 kV, pass through a second, field- free, linear drift region, into the reflector, and finally, to the detector. The detector amplifies and converts the signal to electric current, which is observed and manipulated on a PC-based operating system. For high resolution analysis, the instrument is operated in the reflector mode. After the MALDI process generates the peptide ions, they are accelerated at 20 kV through a grid at 15.6 kV into the first, short, linear, field- free drift region. After this point, the rest of the instrument can be treated as a continuation of this region until the ions enter the reflector and finally reach the detector where, as before, the signal at the detector is amplified and converted to an electrical current.

Samples were prepared by applying 0.6 μl of the sample to a stainless steel 384-well target plate and adding 0.6 μl matrix solution (25 mM α-cyano-4-hydroxycinnamic acid + 10 mM ammonium citrate solution in 50% ACN containing 0.1% TFA). They were allowed to dry at room temperature and were then inserted into the mass spectrometer. All MS and MS/MS experiments were performed twice on the same sample (two spots) and were run in duplicate. Prior to analysis, the mass spectrometer was externally calibrated with a mixture of Angiotensin I, Glu-fibrino-peptide B, ACTH (117), and ACTH (1839). For MS/MS experiments, the instrument was externally calibrated with fragments of Glu-fibrino-peptide.

Interpretation Mascot parameters

For protein identification, a database search was performed, using a local MASCOT server, against the entire NCBI database, to be found at www.ncbi.nlm.nih.gov. Both for PMF and MS/MS searches, the 50 most intense ions were loaded to the server. Search parameters were 50 ppm peptide tolerance and 0.5 Da MS/MS tolerance. CNBr was selected as cleavage reagent and zero miss cleavages were allowed. Conversion of methionine to homoserine lactone or homoserine was allowed as variable modification.

3.2.4 Results and discussion

Optimization of homoserine lactone opening

In order to differentiate the internal CNBr fragments from the C-terminal peptide we developed a chemical approach in which all homoserine lactone derivatives are partially opened (Δm = 18 Da). The complete CNBr mixture was therefore incubated under basic conditions. Parameters explored were (i) the buffer: N,N'-diisopropylamine (DIEA), lutidine (dimethylpiperidine), NH_4HCO_3; (ii) the concentration of the reagent; (iii) the duration of the reaction; and (iv) the reaction temperature. We found that the use of 12.5 mM NH_4HCO_3 (pH 8.0) at 37 °C for 30 min were the best conditions to partially open the homoserine lactone derivatives. Under these conditions, no side reactions were observed for the internal peptides or the C-terminal peptide, therefore the C-terminal peptide appears as the only singlet in the mass spectrum. As the homoserine lactone derivative and the homoserine derivative will probably be detected at different sensitivities, the partial opening was studied in a heuristic fashion. We could also not exclude that a part of the homoserine derivative closes again under the acidic conditions of the MALDI matrix. The partial opening of the homoserine lactone derivative was monitored by MALDI-MS analysis. The relative intensities of both peaks in the doublets varied between 30/70 to 70/30.

Chemical opening in solution

The optimized ring opening was first tested on two model proteins in solution. Therefore, two stock solutions of cytochrome c (horse heart) and alcohol dehydrogenase (yeast) (15 pmol) were prepared and incubated in 12.5 mM NH_4HCO_3 as indicated above. Figure 3.3 A shows the CNBr fingerprint of cytochrome c after chemical opening. The peak at m/z 2779.57 is the C-terminal peptide (with MH_2^{++} at m/z 1390.79) (observed as a singlet) whereas the internal CNBr fragment (Glu66-Met80) appears as a doublet at respectively m/z 1762.92 (homoserine lactone derivative) and m/z 1780.93 (homoserine derivative). The precursor at m/z 2779.57 was subsequently selected for MS/MS analysis, yielding a series of b- and y-ions from which the complete C-terminal sequence could be deduced (Figure 3.3 B). For yeast alcohol dehydrogenase we were also able to differentiate the C-terminal peptide (m/z 1678.91) from two internal peptide fragments (Table 3.1). Interestingly, we also observed a second peptide in the fingerprint as a singlet at m/z 1664.91. The mass difference of 14 Da can be explained by the presence of a C-terminal isopeptide in which residue Ile 338 is replaced by Val as evidenced by MS/MS analysis of both C-terminal peptide fragments (Figure 3.4). This confirms the presence of an isoform with Val at position 338 as observed previously [49].

Table 3.1: Observed CNBr fragments from test and recombinant proteins.

Protein[a]	**Mw (kDa)**	**CNBr fragment**[b]	**Mw calc. (Da)**[c]	**Mw obs. (Da)**[d]
Cytochrome c	11.7	Ile81-Lys104	2779.56	2779.57
Equus caballus		Glu66-Met80	1762.93	1762.92
gi|119388048				
Alcohol dehydrogenase	36.7	Gly76-Met98	2478.22	n.o.
Saccharomyces cerevisiae		Ala169-Met193	2338.22	2338.07
gi|112491285		Glu 333-Lys347	1678.90	1678.78
Elongation factor Tu +	70.0	Leu94-Met128	3955.11	n.o.
N-term GST tag		Phe132-Met365	2627.20	n.o.
Shewanella oneidensis		Asp600-Ala624	2555.43	2555.40
gi|24371827		Pro344-Met365	2532.44	2532.39
		Ala69-Met80	1396.76	1396.75
		Leu81-Met93	1342.67	1342.65
		Val371-Met382	1327.65	n.o
		Asp330-Met343	1297.66	n.o
		Leu154-Met164	1266.65	n.o.
		Val590-Met599	1011.58	n.o.
Glycerophosphodiester	41.7	Tyr336-Lys364	3103.59	3103.73
phosphodiesterase		Val209-Met230	2678.44	2678.55
Haemophilus influenza		Lys2-Met26	2461.30	n.o.
		Ala272-Met287	1693.81	1693.88

[a] NCBI Entrez entries (www.ncbi.nih.gov/Entrez/).
[b] The position of the CNBr fragments in the protein sequence is indicated in Arabic numbers (numbering according to the NCBI entries).
[c] Mw calculated by using the residual monoisotopic values with Met ->hsl (singly protonated) (Only CNBr fragments with a Mw between 1 and 4 kDa are listed).
[d] Mw observed in positive mode reflectron analysis (singly protonated) (n.o.: not observed); C-terminal peptides are indicated in red.

The sensitivity of our method was tested by MS analysis of a dilution series of the cytochrome c stock solution. Therefore, 5, 2.5 and 0.5 pmol cytochrome c were cleaved with CNBr and reacted in 10 μl 12.5 mM NH_4HCO_3. From this solution 0.6 μl was spotted on the MALDI plate and analyzed by MALDI-MS analysis (triplicates). We observed a good fingerprint for the first two fractions, containing respectively 300 and 150 fmol on the plate. No peaks were observed in the fingerprint of the fraction with 0.5 pmol of CNBr digest, containing 30 fmol on the MALDI plate. It should be noted that the sensitivity of our approach will not only depend on the sensitivity of the MS used, but will also depend on the nature (amino acid composition) of the CNBr fragment, its ionization efficiency and suppression effects in complex peptide mixtures. Therefore, to further investigate the sensitivity of our approach, we will, after automation on a TECAN Freedom Evo robot, perform a high-throughput study of 2D-PAGE separated proteins.

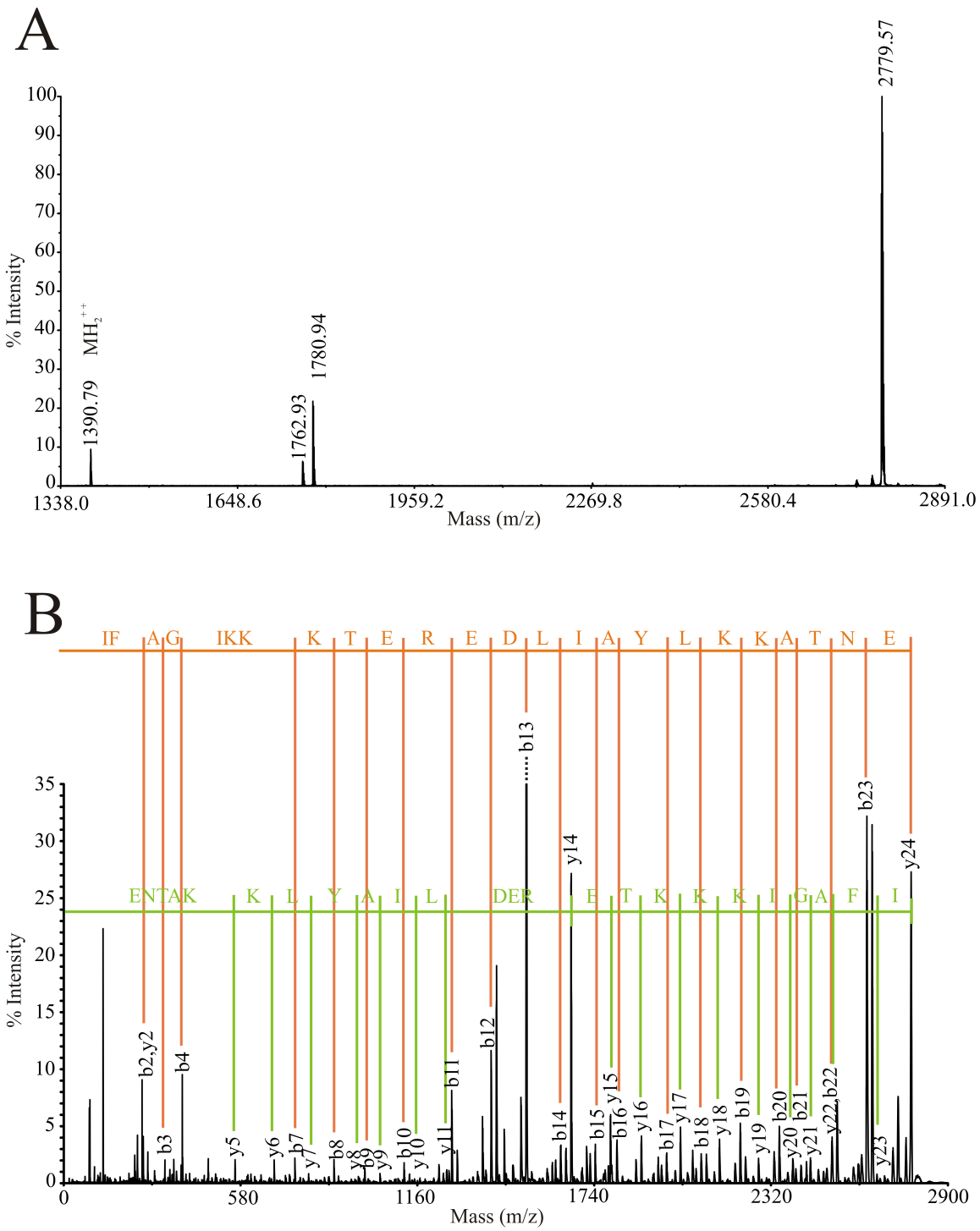


Figure 3.3: C-terminal sequence analysis of horse heart cytochrome c. Panel A: MS spectrum of the CNBr digest after chemical opening. 15 pmol of horse heart cytochrome c was cleaved in solution. 0.6/10 μl of the reaction mixture was applied on the MALDI probe. Panel B: MS/MS spectrum of the C-terminal peptide at m/z 2779.57. Y- and b-ions are indicated in green and red respectively. Amino acid sequence is indicated in the one-letter code.

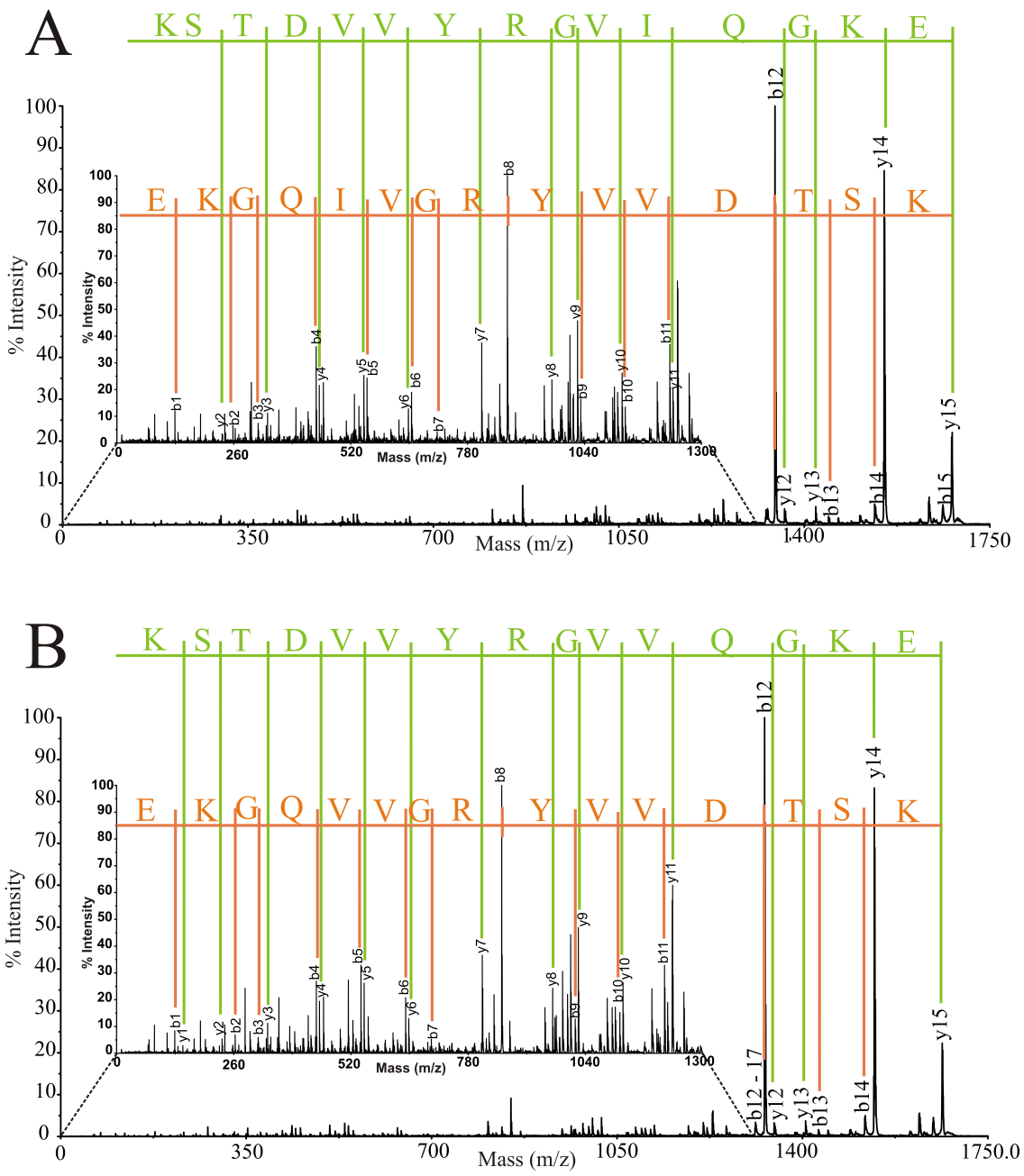


Figure 3.4: MS/MS analysis of C-terminal isopeptides of alcohol dehydrogenase. Panel A: MS/MS spectrum of the C-terminal peptide with m/z 1678.77. Panel B: MS/MS spectrum of the C-terminal peptide with m/z 1664.76 in which Ile338 is replaced by Val. Y- and b-ions are indicated in green and red respectively. Other fragment ions (insets) include a-ions, (a-17)-ions, (b-18)-ions, internal fragment ions and neutral losses of ammonia (internal Arg). The amino acid sequence is indicated in the one-letter code.

As a proof of principle, we applied our method on two recombinant proteins currently under study in our laboratory. In the 70 kDa recombinant TufA protein (EF-Tu), we were able to identify the C-terminal fragment (m/z 2555.44) amongst the internal peptide fragments. Also for the glycerophosphodiester phosphodiesterase, protein D, we were able to differentiate the C-terminal fragment (m/z 3103.73) from the internal fragments (Table 3.1).

Chemical opening of gel purified proteins

To demonstrate the general use of our approach we also applied it to the four proteins separated by SDS-PAGE. 50 pmol of each protein was separated on a precast gel, stained with Coomassie blue and destained. Approximately 1/10 of the gel band was cut (±5 pmol) and cleaved with CNBr in the gel. After extraction and drying of the pooled fraction, the CNBr fragments were opened as indicated above. However, MS analysis (0.6/10 μl) of these fractions indicated the presence of a 25 Da adduct (and multiplexes thereof), both on the C-terminal fragment and the partially opened internal fragments, as observed in some of the previous experiments. Although the exact nature of this modification remains unclear so far, we observed that the reaction occurs only after treatment in the slightly basic buffer, as no adducts were observed in the untreated mixtures. Furthermore, if the CNBr peptide mixture was first purified by an additional ZipTip extraction (C18) before the chemical opening, no adducts were observed.

The success of this approach was demonstrated with the recombinant protein D. After ZipTip and chemical opening, four peptides were observed of which only one appeared as a singlet (Figure 3.5 A). MS/MS analysis of the precursor at m/z 3103.73 yielded the complete C-terminal sequence of 29 amino acids (b- and y-ions) (Figure 3.5 B). However, MS/MS analysis of the C-terminal peptide from the EF-Tu yielded no good fragmentation spectrum. This is consistent with our previous observations that some C-terminal CNBr fragments do not fragment well (unpublished results). It should be noted that CNBr fragments do not contain a C-terminal Lys or Arg and that fragmentation will depend on the presence and location of internal basic residues, and on the sequence and the amino acid composition of the peptide itself. To the best of our knowledge, this problem has not been studied in detail so far.

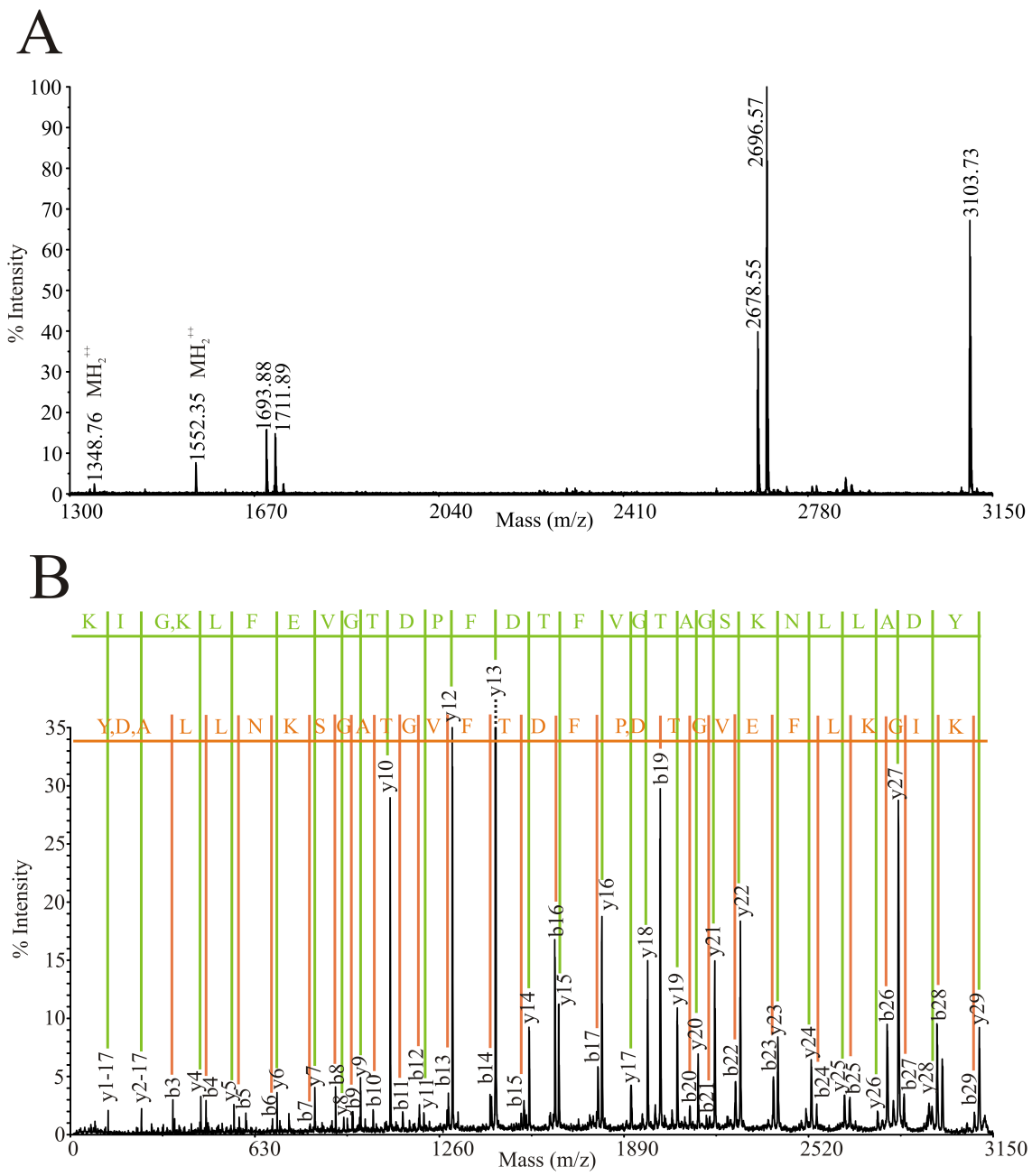


Figure 3.5: C-terminal sequence analysis of recombinant protein D. Panel A: MS spectrum of the CNBr digest after chemical opening; 50 pmol of recombinant protein D was separated by SDS-PAGE. Approximately 1/10 of the band was cut out of the gel and cleaved by CNBr. Panel B: MS/MS spectrum of the C-terminal peptide at m/z 3103.73. Y- and b-ions are indicated in green and red respectively. The amino acid sequence is indicated in the one-letter code.

3.2.5 Conclusion

Shotgun proteomics provides the most comprehensive identification of proteins from cellular lysates, but generally fails to characterize the N- or C-terminal sequence of the protein under study. This is most often due to the scare detection of terminal peptides during the mass spectrometric analysis of complex peptide mixtures and the incomplete annotation of protein termini in protein databases. Here, we have demonstrated a novel approach allowing to discriminate C-terminal peptides in CNBr mixtures. The method can be applied at the low femtomol level and can be used for the analysis of gel or gel-free purified proteins. For gel-separated proteins we observed an adduct of 25 Da during chemical opening. Although its exact nature remains unknown so far, this adduct is not observed if the gel fractions are desalted (ZipTip) before the ring opening. Although some peptides had a molecular mass in the good range for MS analysis (1-4 kDa) they were not observed during mass analysis (Table 3.1). This is most likely due to a low ionization efficiency (depending on the sequence and amino acid composition) or to sample suppression effects as observed previously[42, 43].

In contrast to the carboxypeptidase method, the chemical approach is suitable for the development of a high-throughput approach, by coupling the MALDI-MS/MS analysis to a robotic sample preparation device. For that purpose a TECAN Freedom EVO robot has recently been introduced in our laboratory. The development of a method that allows a high-throughput analysis of proteolytic processing at a proteomics level will be of fundamental importance to gain a better understanding of the complex biological processes in living organisms.

3.2.6 Acknowledgements

B.S. is a Postdoctoral fellow of the Fund for Scientific Research-Flanders (F.W.O.-Vlaanderen, Belgium). P.M. is funded by a Ph.D. grant of the Institute for the promotion of Innovation through Science and Technology in Flanders (I.W.T.-Vlaanderen). The authors acknowledge funding by the Fund for Scientific Research-Flanders through Research Grant G.0644.07 (F.W.O.-Vlaanderen). The authors would like to thank Ester Behiels for the kind gift of recombinant proteins.

3.3 Automation of C-terminal sequence analysis of 2D-PAGE separated proteins.

P.P. Moerman[1], K. Sergeant[2], G. Debyser[1], I. Timperman[1], B. Devreese[1]*, B. Samyn[1]

[1] Laboratory for Protein Biochemistry and Biomolecular Engineering, Ghent University, Ghent, Belgium
[2] Department Environment and Agro-biotechnologies, Centre de Recherche Public, Gabriel Lippmann, Belvaux, Luxembourg

*Corresponding author: Bart.devreese@ugent.be

This chapter has been published in EuPA Open Proteomics (2014), 3, 250-261

3.3.1 Abstract

Experimental assignment of the protein termini remains essential to define the functional protein structure. Here, we report on the improvement of a proteomic C-terminal sequence analysis method. The approach aims to discriminate the C-terminal peptide in a CNBr-digest where Met-Xxx peptide bonds are cleaved in internal peptides ending at a homoserine lactone (hsl)-derivative. pH-dependent partial opening of the lactone ring results in the formation of doublets for all internal peptides. C-terminal peptides are distinguished as singlet peaks by MALDI-TOF MS and MS/MS is then used for their identification. We present a fully automated protocol established on a robotic liquid-handling station.

3.3.2 Introduction

Characterization of the exact N- or C-terminus of a protein is an essential contribution to evidence-based gene annotation. Current prediction methods fail to provide a proper assignment of these termini that are highly sensitive to post-translational processing events. The failure of classical chemical protein sequencing methods to reach the throughput and sensitivity to routinely provide this information forced us to adopt proteomic strategies typically involving isolation of terminal peptides. Replacement of chemical protein sequencing by top-down mass spectrometry is considered, but requires costly high resolution (FT-based) instrumentation and intensive protein purification. Therefore, full exploitation of this approach awaits further technical and bioinformatics improvements [50, 51].

A number of methods have been described for the specific isolation of the C-terminal peptide of a protein, basically falling into three categories, i.e. introducing labels at certain functional groups, chromatographic enrichment of terminal peptides and binding via functional groups to a resin. The introduction of an isotopic label, either in the C-terminal peptide using Br-containing reagents [52] or in the internal peptides using $H_2{}^{18}O$ during in-gel digestion [53], allows to distinguish the C-terminal peptide by matrix-assisted laser desorption/ionization (MALDI)-MS. Due to the difficulties to selectively target the C-terminus and/or the need for a large amount of sample, these techniques have rarely been applied [54]. In a second strategy, N-terminally blocked peptides and C-terminal peptides are enriched from a complex peptide mixture by strong cation exchange (SCX) purification at low pH [48]. An example of the third strategy is the use of immobilized anhydrotrypsin to specifically bind tryptic peptides that contain an arginine or lysine at their C-terminus. C-terminal peptides, typically devoid of such a C-terminal basic residue, are then recovered from the non-bound fraction [55]. Lastly, ProC-TEL uses a positive selection approach to enrich the C-terminal peptides using the transpeptidase activity of carboxypeptidase Y to label the C-terminus of the protein with biotin prior to tryptic digestion [56].

Recently two C-terminal sequencing techniques have been reported that can be applied to complex biological samples, both combining two of the previously mentioned strategies. Based on the COFRADIC approach, N-terminally blocked and C-terminal peptides are first enriched in a SCX step followed by the introduction of a butyrate label on the α-amine of the C-terminal peptides [57]. This allows identifying the latter by their altered chromatographic behavior. The C-TAILS approach combines multiple amine and carboxyl group protection steps with a tryptic digest to selectively bind the internal peptides to a primary amine containing resin [58]. These methods have successfully been applied to complex samples, but require multiple derivatization and/or separation steps making them labor intensive.

In the so-called ladder sequencing techniques, first described by Chait *et al.* [59], a sequence-defining concatenated set of peptide fragments, each differing from the next by a single residue, is generated in a controlled fashion. Subsequently the complete fragment set, the peptide ladder, is analyzed mostly by MALDI-MS. While chemical ladder generating procedures, often based on the Edman degradation, have mainly been developed for N-terminal approaches, proteolytic digestion using carboxypeptidases (CPase) has been the preferred method for C-terminal sequence analysis. Our group has developed an enzymatic ladder sequencing technique that enables the systematic identification of C-termini from proteins either in-gel or in-solution [42, 43]. The proteins are first chemically cleaved under acidic conditions with cyanogen bromide (CNBr). CNBr hydrolyses the peptide bond C-terminal from methionine residues that are converted into homoserine (hse). Under acidic conditions hse undergoes a cyclisation to the lactone form [1]. This homoserine lactone (hsl) residue allows to differentiate between the C-terminal peptides and the N-terminal and internal peptides. During incubation with a CPase, only the original C-terminal fragment (having a free carboxyl group) is accessible to enzymatic degradation and forms a ladder. The sequence of the C-terminal peptide can then be read by measuring the unseparated peptide mixture containing the CPase generated fragments on a MALDI-TOF MS (Figure 3.6).

In this ladder sequencing technique, CPY and CPP are selected because of their broad amino acid specificity [44]. However, it is known that cleavage C-terminal of Gly is slow, and also the presence of Phe, Thr, or Lys slows down or inhibits ladder generation [60]. Additionally, the rate of hydrolysis depends on reaction conditions such as pH, ionic strength and substrate concentration. Due to these limitations, we observed that the reaction times required optimization for each individual sample. Finally, it should be noted that most exopeptidases have a K_m-value in the range of 5-50 μM, which means that they are operating at 50% maximum velocity when a protein concentration of 5 pmol/μl is used [61]. At lower concentrations (sub pmol), the proteolytic activity toward the substrate will be minimal.

We recently reported the modification of the technique by eliminating the use of CPase through developing a novel chemical approach to differentiate between internal and the C-terminal peptides without the need to separate the peptides prior to analysis [60]. In this chemical approach, the peptide mixture, generated by digestion with CNBr, is incubated in slightly basic buffer. This results in a partial opening of the hsl derivatives to the corresponding hse derivative (Δm = +18 Da). All internal peptides appear as doublets in the MALDI-TOF MS spectrum, whereas the C-terminal peptide is the only singlet present and can be selected for MS/MS analysis (Figure 3.6). By replacing the CPase-dependent determination of the C-terminal peptide by a chemical method, we made the technique sequence independent and eliminated the need to optimize the protocol for each sample separately.

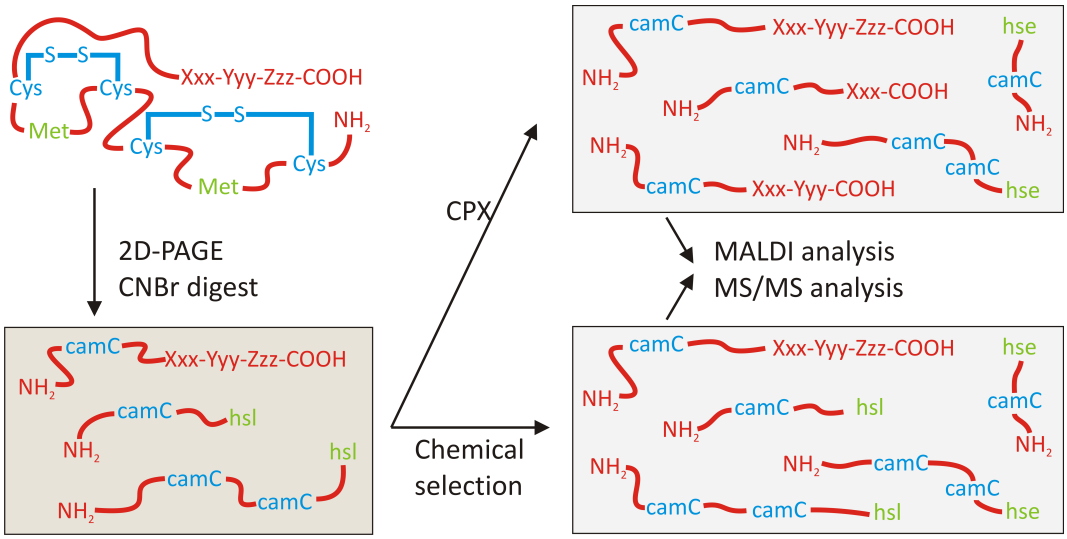


Figure 3.6: Schematic representation of the different steps in the two C-terminal sequencing methods. Proteins separated by 2D-PAGE are cleaved in gel with CNBr after destaining. CNBr cleavage results in the formation of internal fragments ending at a homoserine lactone (hsl) derivative. When the fragments are incubated with carboxypeptidase (CPase), only the peptide containing the original C-terminal sequence (Xxx-Yyy-Zzz) is accessible for enzymatic degradation by CPase and forms a ladder. When chemical selection is used to differentiate internal from C-terminal peptides, the peptides are incubated in a slightly basic buffer resulting in the partially opening of the hsl ring forming both homoserine (hse) and hsl derivatives of the internal peptides. Hsl and hse have a mass difference of 18 Da. Both the ladders and the hsl/hse derivatives are analyzed by MALDI analysis on a 4800 MALDI TOF/TOF instrument. CPX represents a CPase or a mixture of CPases.

While our previous work was based on some model proteins in-gel and in-solution, we here report the results obtained from a proof-of-concept experiment by performing the enhanced protocol in a proteomics setup on 2D-PAGE gel separated proteins from *Shewanella oneidensis* MR-1. To increase the throughput of the method, we transferred it to a liquid handling and robotic platform. After loading the protein gel spots in a 96-well plate, the robot performs all necessary steps of the protocol in an automated way, including sonication, heating, cooling, pipetting and

sample clean- up using ZipTips. Multi-arm robots can move the 96-well plates between different modules on the worktable limiting manual interference and contaminations. We compared the results with data obtained by C-terminal sequencing using the manual CPase ladder sequencing approach and demonstrate a significant improvement of the number of identified C-terminal sequences.

3.3.3 Materials and methods

Materials and chemicals

HPLC-grade acetonitrile (ACN) was obtained from BioSolve (Valkenswaard, The Netherlands). Trifluoroacetic acid (TFA) (purity >99.9%) was obtained from Beckman Instruments (Palo Alto, CA, U.S.A.). Urea was obtained from GE Healthcare (Diegem, Belgium). 'Complete mini' ethylenediaminetetraacetic acid (EDTA)-free protease inhibitor mix, DNase I and RNase came from Roche (Vilvoorde, Belgium). Coomassie Plus Bradford protein assay was purchased from Thermo (San Jose, CA, USA). ReadyStrip IPG strips, Bio-Lyte 3-10 ampholyte and Tris/Glycine/SDS running buffer 10x were purchased from Bio-Rad (Hercules, CA, US). The solution of 30% (w/v) acrylamide/0.8% (w/v) bisacrylamide was purchased from National Diagnostics (Atlanta, GE, US), whereas agarose was from Eurogentec (Liege, Belgium). Sodium dodecyl sulfate (SDS) was obtained from Merck (Darmstadt, Germany). Sequencing-grade CPY was obtained from Roche (Indianapolis, IN, USA). C-18 Zip-Tips were obtained from Millipore (Billerica, MA, USA). 2 ml glass vials with screw top, PTFE/Red rubber septa and polypropylene screw caps were obtained from Supelco (Bellefonte, PA, US). Polystyrene V-shaped 96-well plates were obtained from Greiner Bio-one (Frickenhausen, DE). Water was purified using a MilliQ water filtration system (Millipore). Other chemicals and reagents were purchased from Sigma (St. Louis, MO, US).

Bacterial growth and protein extraction

S. oneidensis MR-1 was grown aerobically overnight in 1 L Luria Bertani (LB) medium on a rotary shaker at a speed of 220 rpm at 28 °C until an optical density at 600 nm (OD600) of 0.6 was reached. The cells were then centrifuged and washed twice using a 50 mM TrisHCl solution (pH 7.5). The bacterial cell pellet was dissolved in lysis buffer pH 7.5 (9 M urea, 40 mM Tris-HCl, 2% 3-[(cholamidopropyl) dimethylammonio]- 1-propanesulfonate (CHAPS), 1% dithiothreitol (DTT), 0.5 mg/ml bovine pancreas DNase I, 0.25 mg/ml bovine pancreas RNase A, 50 mM $MgCl_2$) containing a protease inhibitor mixture (EDTA-free). A volume of 1.5ml of this solution was added per gram of biomass and sonicated on ice using a Digital Sonifier S-250D (Branson, Danbury, CT) for 1 min on 30% amplitude in pulses of 2 s. After sonication the sample was kept on ice for 15 min to allow the DNase and RNase to digest the polynucleotides. Next the sample was centrifuged at 16,000 g for 45 min and the soluble

protein fraction was precipitated with acetone at 20 °C. The protein pellet was dissolved in rehydration buffer (9 M urea, 1% DTT, 2% CHAPS, 2% Bio-Lyte 3-10 ampholyte solution).

2D-PAGE

350 μl of bacterial extract ($\pm$300 μg of protein as determined by a Bradford protein assay) was loaded via passive in-gel rehydration [62] on a 17 cm IPG strip, pH range 4-7. The strips were covered with mineral oil and were left for 9 h at room temperature. Isoelectric focusing (IEF) was performed using a Protean IEF cell (Bio-Rad, Hercules, CA) at room temperature by applying a stepwise voltage gradient up to 3500 Volt, until 35 kVh were reached. Following the focusing step the proteins were reduced in 50 mM Tris-HCl (pH 8.8), 6 M urea, 2% SDS, 30% glycerol, 1% DTT) and amidoacetylated (same buffer, DTT replaced by 5% Iodoacetamide (IAA)) by shaking the strip submerged in buffer for 10 min at room temperature. For the second dimension, the strips were placed on a 12% SDS-PAGE gel with a 0.5% agarose gel as interface. The gels were run at 30 mA/gel until the bromophenol blue front reached the bottom of the gel. After fixation (40% EtOH, 10% acetic acid), the gels were stained overnight with Coomassie G. The gels were destained with 30% methanol prior to spot picking.

CNBr cleavage

CNBr cleavage of gel separated proteins was performed as described before [43, 60]. Briefly, after visualization and destaining, the bands or spots containing the protein were excised from the gels. Before cleavage with CNBr, the gel pieces were washed twice with 100 μl 200 mM NH_4HCO_3 /50% ACN and then shrunken with 40 μl ACN and rehydrated in 5 μl MQ. CNBr cleavage was started by adding 15 μl TFA and 5 μl 5 M CNBr in ACN. CNBr and TFA are highly toxic and corrosive products which must, at any time, be manipulated under a fume hood, only by skilled personnel wearing protective clothing! After incubation overnight (4 °C) the supernatant was collected and the peptides were extracted twice with 30 μl 70% ACN/0.1% TFA for 30 min at 37 °C. All fractions were pooled and dried in a SpeedVac (Thermo Savant, San Jose, CA, USA). Care must be taken during sample manipulation to avoid oxidation. Oxidation of methionine to methionine sulfoxide prevents CNBr from attaching to the sulfur atom and initiating the cleavage reaction. Therefore, samples were submitted for gel electrophoresis as soon as possible after cell lysis, and gels were kept in a tris(2-carboxyethyl)phosphine (TCEP) containing solution until CNBr digestions.

Carboxypeptidase protocol

The CPase protocol was performed as described by Samyn *et al.* [43]. Sequencing-grade CPY was diluted to a stock solution of 1 pmol/μl in 40 mM sodium citrate pH 6.0. For time-dependent ladder formation the CNBr fragments were dissolved in 10 μl 10 mM ammonium acetate buffer

pH 5.4 and mixed with CPY (enzyme to substrate ratio of 1:50 wt/wt). After 0, 1, 3, 10, 20 and 30 min 0.5 μl of this mixture and 0.5 μl of matrix were spotted on the MALDI target plate. For concentration-dependent digestions, the CNBr fragments were dissolved in 5 μl ammonium acetate buffer. 0.5 μl was spotted onto the MALDI target plate and incubated with 0.5 μl of 1, 0.2, 0.04, 0.008 pmol/μl CPY until solvent evaporation terminated the reaction.

Chemical derivatization

Before the chemical derivatization, the dried samples were desalted using C18 ZipTip micro purification tips. The ZipTip protocol was performed as described by the manufacturer. A 50% ACN/0.1% TFA solution was used as activation and elution solvent. A 0.1% TFA solution in MQ was used as dissolving, equilibration and wash solvent. After elution the samples were dried (in an oven) at 60 °C for 15 min. The partial opening of the hsl to hse was performed by dissolving the sample in 10 μl 10 mM NH_4HCO_3 (pH 8.0), and incubation at 37 °C for 30 min. During the dissolving steps the samples were sonicated for 5 min. After incubation 0.5 μl of the sample was spotted onto the MALDI target plate

Automation

All samples obtained from the chemical derivatization protocol were processed on a Tecan Freedom Evo 150 platform (Figure 3.7). The custom designed robot is equipped with a robotic manipulator arm to move 96-well plates and an 8-channel liquid handling arm. Two channels on the liquid handling arm are set up to be used with ZipTips. Two other channels are equipped with high internal volume stainless steel needles with sharp tips, used to pierce through the rubber septa of the vials containing the corrosive chemicals. The last 4 channels on the liquid handling arm are equipped with high precision Tecan Positioning System (Te-PS) fluorinated ethylene propylene (FEP) coated stainless steel needles used for all other liquid handling. The worktable (1.5 m) holds 96-well plate carriers, solvent trays, slots for glass vials, a heater with six 96-well plate slots at different temperatures and a sonicator used as alternative for a vortex. The solvents and one 96-well plate carrier were cooled to 4 degrees to slow down evaporation of the solvents. The entire robotic platform is placed under a fume hood to protect the user from the toxic TFA and CNBr vapors. Freedom Evoware 2 standard software serves as an interface between the user and the platform.

The excised gel spots were collected in polystyrene V-shaped 96-well plates that are resistant to the corrosive properties of CNBr. All steps of the chemical derivatization protocol; destaining, chemical cleavage, extraction, ZipTip, lactone opening and spotting on the MALDI target plate were performed by the robotic platform. The only steps that required manual handling were covering the 96-well plate with Parafilm to prevent evaporation during the overnight incubation with CNBr and placing the 96-well plate in the SpeedVac to dry the samples after extraction.

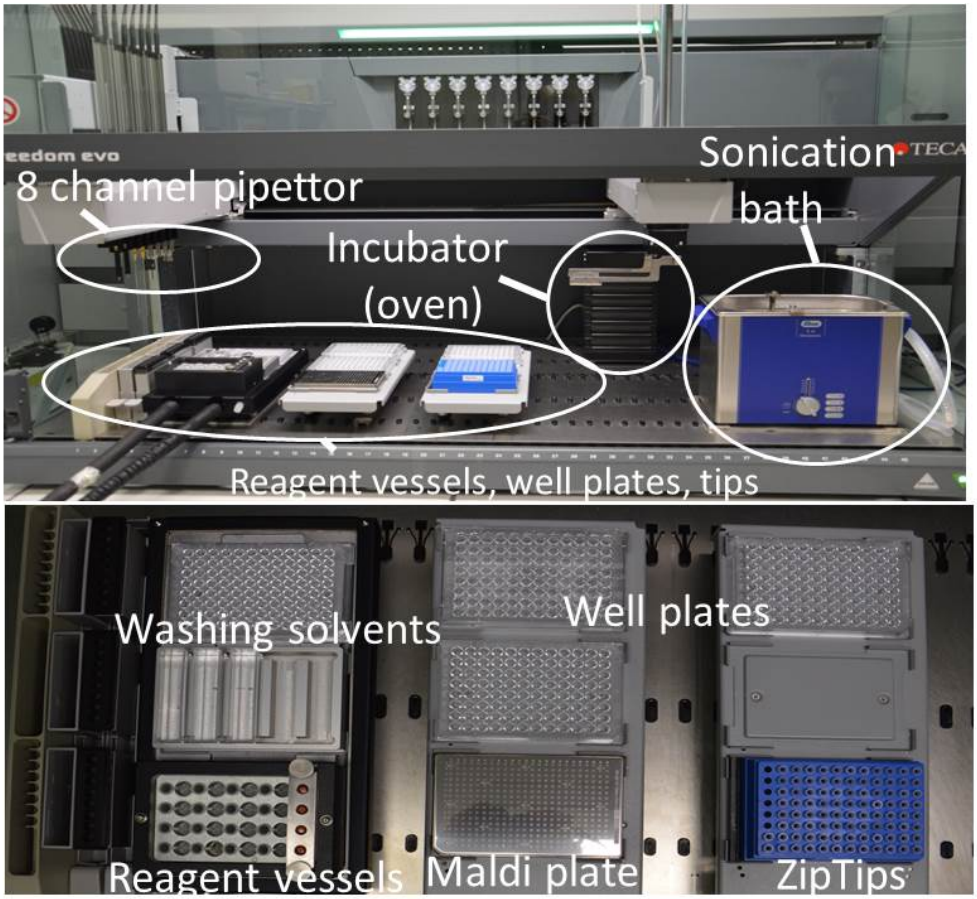


Figure 3.7: Picture of the setup of the Tecan liquid handling and robotic system. Top: front view, Bottom: top view on the reaction table.

Mass spectrometry

All mass spectrometry analysis were performed on an Applied Biosystems 4800 plus Proteomics Analyzer with TOF/TOF optics (Applied Biosystems, Foster City, CA). This MALDI mass spectrometer uses a 200 Hz frequency tripled Nd:YAG laser operating at a wavelength of 355 nm. For high resolution analysis of the peptides, the instrument was operated in reflector mode. For MS/MS, ions generated by the MALDI process were accelerated at 8 kV through a grid at 7.3 kV into a short, linear, field-free drift region. In this region, the ions passed through a timed-ion-selector device that is able to select a peptide ion, for subsequent fragmentation in the collision cell. The selected ions then passed through a retarding lens where they were decelerated and allowed to enter into the collision cell, which was operated at 7 kV. The collision energy is defined by the potential difference between the source and the collision cell (1 kV). After passing through the collision cell, the ions (both intact peptide ions and fragments) were accelerated in the second source region at 15 kV into the reflector, and finally, to the detector.

Samples were prepared by applying 0.5 μl of the sample to a 384-well stainless steel target plate and by adding 0.5 μl matrix solution (25 mM α-cyano-4-hydroxycinnamic acid + 10 mM ammonium citrate solution in 50% ACN containing 0.1% TFA). They were allowed to air-dry at room temperature and were then inserted into the mass spectrometer and subjected to MALDI-MS analysis. All MS and MS/MS experiments were performed twice on the same sample (two spots) and were run in duplicate. Prior to analysis, the mass spectrometer was externally calibrated with a mixture of Angiotensin I, Glu-fibrino-peptide B, ACTH (1-17), and ACTH (18-39). For MS/MS experiments, the instrument was externally calibrated with fragments of Glu-fibrinopeptide.

MS data analysis

GPS explorer (Applied Biosystems, Foster City, CA) was used to search the combined MS and MS/MS data against the NCBI *S. oneidensis* MR-1 protein database [29], obtained from www.ncbi.nih.gov/Entrez/. GPS Explorer used a local MASCOT server for protein identifications [63]. Both for PMF and MS/MS searches, the 50 most intense ions were uploaded to the server. Search parameters were as follows: 100 ppm peptide tolerance and 0.5 Da MS/MS tolerance. CNBr was selected as cleavage reagent and zero miss cleavages were allowed. Conversion of methionine to hsl or hse and carbamidomethylation of cysteine were allowed as variable modifications.

The MS/MS spectra of the C-terminal peptides were manually interpreted and the sequence tags were searched against the NCBInr database using BLASTp (http://blast.ncbi.nlm.nih.gov/). *S. oneidensis* MR-1 was selected as organism and BLOSUM 62 was selected as scoring matrix. Additionally, to confirm the identifications, the obtained C-terminal sequence tags were searched using MS homology (http://prospector.ucsf.edu/prospector/mshome.htm).

3.3.4 Results

Implementing the chemical derivatization protocol on a robotic platform

After optimization of our chemical method [60] we observed that it was robust and, in contrast with our previous CPase-based strategy, it was largely independent of the sequence of the C-terminal peptides [42]. In order to obtain a higher throughput, we decided to transfer the protocol to an automated robotic platform. Due to some limitations of the Tecan platform and the corrosive properties of two essential chemicals (TFA and CNBr) some adaptations had to be made to the initial protocol (The detailed protocol can be found online at doi:10.1016/j.euprot.2014.03.004).

To limit the number of manual interventions, all drying and dissolving steps were performed in an oven (60 °C) and in a sonicator embedded in the robot instead of being placed manually in a SpeedVac and vortex, respectively (Figure 3.7). Only the peptide extracts obtained after CNBr digestion were dried using the SpeedVac because of their large volume.

To protect the user and the electronics against exposure to toxic chemicals and to prevent evaporation of the solvents and the samples during chemical degradation and incubation at elevated temperature steps, the samples and solvents were kept cooled and sealed as much as possible. Because the CNBr/ACN and TFA solutions degrade plastic solvent containers, they were kept in cooled glass vials with rubber septa that can be pierced with inert needles. The tip of the needles was cut to a 45° angle to allow them to easily pass through the thick septa. The needles also have a large internal volume to prevent the CNBr and TFA containing solutions from entering and degrading the plastic tubing of the liquid handling system. In view of the observation that the CNBr and TFA containing solutions degraded most of the tested glue based seals for 96-well plates, Parafilm was found to be the best alternative to avoid evaporation during overnight chemical digestion.

Next to the vials, rubber caps, needles and seals, a large set of 96-well plates were tested for compatibility with the solvents and recovery of peptides after drying steps. During all these tests SDS-PAGE separated horse heart cytochrome c was used as a test protein. The Polystyrene V-shaped 96-well plates from Greiner Bio-One were finally selected as they provided the most reproducible results (data not shown).

Evaluation of the automated procedure on *Shewanella oneidensis* MR-1 2D-PAGE gel spots

The automated method was evaluated for proteome analysis of *S. oneidensis* MR-1. *S. oneidensis* was aerobically grown and the total protein extract was separated by 2D-PAGE. After Coomassie blue staining, 96 of the most intense spots were selected (Figure 3.8) and loaded in a 96-well plate. The robot autonomously performed all steps from the sample preparation protocol, starting from destaining the spots and finishing by spotting the peptides on the MALDI-plate, in 24 h with only 2 minor manual interventions (applying Parafilm and transferring the 96-well plate to the SpeedVac).

To illustrate our approach, we focus on the CNBr peptide map of spot 21. After incubation with the basic buffer, an internal peptide appears as a doublet (2339.0 Da and 2357.0 Da) with a mass difference of 18 Da in the spectrum. The C-terminal peptide appears as the most intense singlet peak at 2089.9 Da (Figure 3.9 A). After fragmentation, a complementary b- and y-ion series was obtained covering the entire C-terminal peptide sequence (Figure 3.9 B). The

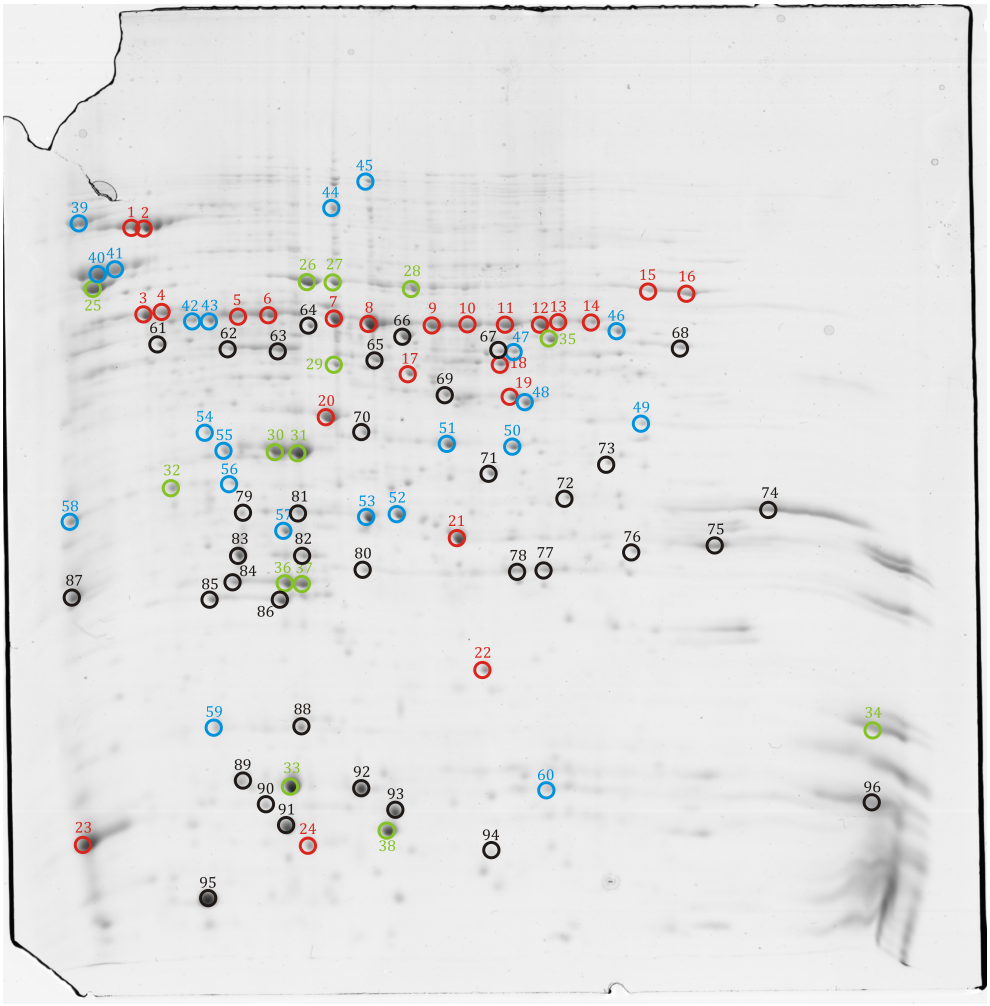


Figure 3.8: 2D-PAGE separated proteins from aerobically grown *Shewanella oneidensis* MR-1. ±300 μg of total cell extract was loaded on an IPG strip (4-7) and analyzed as described. The numbered spots were subjected to both approaches. The spots are coloured according to the C-terminal sequence information and protein identification obtained during chemical selection. The 24 spots marked in red were identified only using the C-terminal sequence obtained. For the 14 spots marked in green, the C-terminal peptide was identified, but no sequence information was obtained. No C-terminal sequence information was obtained for the 22 spots marked in blue, but the proteins were identified by PMF analysis. The 36 spots marked in black were not identified.

18-residue amino acid sequence was submitted to a BLAST search against the *S. oneidensis* proteins in the NCBI database [29, 64] and was identified as the C-terminal peptide of a protein annotated as an ABC-type tungstate uptake system substrate-binding component (TupA).

Out of 96 samples, 44 unique proteins from 60 different spots, with a molecular mass ranging from 12.5 to 93 kDa, could be identified using the CNBr peptide mass peptide fingerprint and MS/MS data (Table 3.2 and Appendix Table A.1). In 38 spots (24 proteins) thereof, we were able to distinguish the C-terminal peptides since they appeared as singlets in the MS

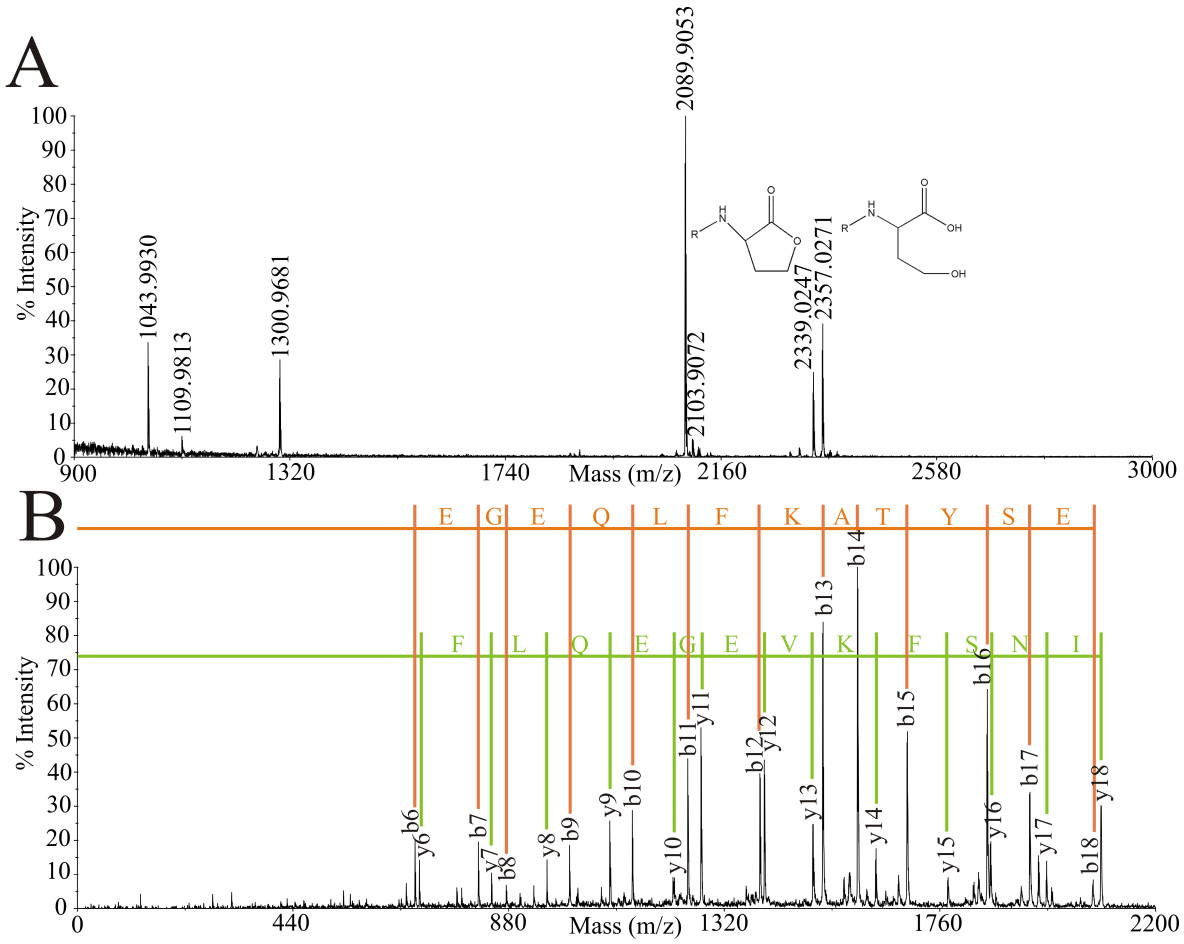


Figure 3.9: C-terminal sequence analysis of spot 21 (ABC-type tungstate uptake system substrate-binding component TupA) using chemical selection. Panel A: MS spectrum of the CNBr digest after chemical opening. Masses with m/z 1043, 1109 and 1300 are reoccurring contaminants. Panel B: MS/MS spectrum of the C-terminal peptide at m/z 2089.9. Y- and b-ions are indicated in green and red respectively. Amino acid sequence is indicated in one-letter code.

spectrum (Table 3.2). Other singlet peaks were recurrent contaminants that could be excluded. These C-terminal peptides were selected for MS/MS analysis and after manual interpretation of the obtained fragmentation spectra we were able to obtain sufficient C-terminal sequence information to identify 13 proteins present in 24 spots using a BLAST search against the *S. oneidensis* MR-1 protein database (Table 3.2). In 4 cases, the entire sequence of the C-terminal peptide could be extracted from the fragmentation spectrum while in the other, a sequence tag of minimum 6 consecutive amino acids could be determined. In one case, DNA-binding protein H-NS family (spot 38), the C-terminal peptide contained one missed cleavage due to the presence of an oxidized methionine.

Table 3.2: Automated C-terminal sequence analysis of 2D PAGE-separated proteins of *Shewanella oneidensis* MR-1

Protein[a]	**Spot**[b]	**Mw (kDa) /pI**	**Mw calc. (Da)**[c]	**Mw obs. (Da)**[d]	**Sequence C-term peptide**[e]	**blastp score**[f i]	**MS hom. score**[g i]	**GPS score (>49)**[h i]
Ribosomal protein S1 RpsA gi|24373949	1,2	61.2/4.91	990.56	991.53	AE**AFKAARK**	5,0E-03	33	M
ATP synthase F1 β-subunit AtpD gi|24376219	3,4	49.7/4.88	1614.89	1615.91	VGSI**DEAVEKANKKK**	6,0E-07	54	79
Translation elongation factor Tu TufA gi|24371827	5-14	43.3/5.08	2554.43	2555.40	**DE**GL**R**FAI**REGGRT**VGAG-VVAKIIA	1,3E-02	72	82
Translation elongation factor Tu TufB gi|24371815	5-14	43.3/5.13	2588.42	2589.37	**DEGLRFAIREGGR**TVGAG-VVAKIFA	3,0E-09	67	65
Inosine-5'-monophosphate dehydrogenase GuaB gi|24374804	15,16	51.6/6.45	2283.09	2283.96	GE**SHVHDVTITKEAPN**YR-SGS	2,0E-10	75	M
Isovaleryl-CoA dehydrogenase LiuA gi|24347753	17	42.0/5.56	1320.70	1321.69	LIGR**ELYNESK**	1,0E-03	36	144
Leucine dehydrogenase Ldh gi|24374179	18	37.1/5.77	1061.60	1062.50	**ARAIYQAAKA**	7,0E-06	46	M
Fructose-bisphosphate aldolase, class II Calvin cycle subtype Fba gi|24346525	19	38.5/5.71	1679.89	1680.79	**YKAYQSGALDPKINL**	2,0E-11	78	63
NAD dependent malate dehydrogenase Mdh gi|24372359	20	32.1/5.37	1759.97	1760.84	**LDTLKGDIKLGVDFVK**	6,0E-12	80	75
ABC-type tungstate uptake system substrate-binding component TupA gi|24376191	21	29.3/6.46	2088.80	2089.91	**INSFKVEGEQLFKATYSE**	1,0E-14	90	275
Transcription antitermination protein NusG gi|24371817	22	20.9/5.74	2008.03	2009.05	IFGR**STPVELDF**SQVE**KG**	2,0E-04	x/41	96

continued on next page

Table 3.2: Automated C-terminal sequence analysis of 2D PAGE-separated proteins of *Shewanella oneidensis* MR-1. – *continued from previous page*

Protein[a]	**Spot**[b]	**Mw (kDa) /pI**	**Mw calc. (Da)**[c]	**Mw obs. (Da)**[d]	**Sequence C-term peptide**[e]	**blastp score**[f i]	**MS hom. score**[g i]	**GPS score (>49)**[h i]
50S ribosomal protein L7/L12 RplL gi|24371821	23	12.5/4.63	3494.87	3495.30	SE**AAPVAVKEGVSKEE**AE-**ALKKELVEAGASVEIK**	4,0E-15	156/147/72	87
ATP synthase F1 ϵ-subunit AtpC gi|24376218	24	15.1/5.35	1738.06	1738.91	AQLR**VVETIKK**NIAR	3,0E-03	32	M
Trigger factor peptidyl-prolyl cis-trans isomerase Tig gi|24373359	25	47.6/4.87	716.39	717.39	NKATGRA			107
ATP synthase F1 α-subunit AtpA gi|24376221	26,27	55.1/5.38	4204.10	4204.90	NSEHAALIKLINETGDYNA-DIEAELKAGLDKFVATQTWox			78
Dihydrolipoamide dehydro-genase LpdA gi|24372021	28	50.5/5.56	4610.36	4611.10	GCDAEDLALTIHAHPTLHE-SVGLAAEIYEGSITDLPNP-KAKKK			M
bifunctional acetylornithine aminotransferase/succinyl-diaminopimelate aminotrans-ferase/succinylornithine transaminase ArgD gi|24346120	29	43.2/5.43	3621.96	3623.32	AGANVVRFAPSLVIPEADI-AEGLARFERAVASIAAA			M
Elongation factor EF-Ts gi|24373198	30,31	30.4/5.31	5430.94	5432.70	EPKKTVGEFLKEKGAKVTN-FIRLEVGEGIEKKEEDFAA-EVAAQIAASKKA			108
Translation elongation factor P Efp gi|24373875	32	20.6/4.79	3883.11	3883.46	KPATITGGGTISVADFVKVG-DKIEIDTRTGEFKKRV			M
30S ribosomal protein S6 RpsF gi|24375418	33	15.0/5.26	3205.50	3206.29	AKAKDERDSRRGPAGDRSY-DEANAEEIAE			122
50S ribosomal protein L5 RplE gi|24371841	34	20.2/9.27	3025.58	3026.18	DIVITTSAKTDEEGRALLDAF-NFPFKK			135
Alanine dehydrogenase Ald gi|414562022	35	39.1/5.89	2695.44	2696.39	HGKLVCKEVAQALNLEYTA-PTGLLA			79

continued on next page

Table 3.2: Automated C-terminal sequence analysis of 2D PAGE-separated proteins of *Shewanella oneidensis* MR-1. – *continued from previous page*

Protein[a]	**Spot**[b]	**Mw (kDa) /pI**	**Mw calc. (Da)**[c]	**Mw obs. (Da)**[d]	**Sequence C-term peptide**[e]	**blastp score**[f i]	**MS hom. score**[g i]	**GPS score (>49)**[h i]
Peroxiredoxin TsaA gi|414562066	36,37	21.8/5.19	1708.83	1709.71	NASTAGVAAYLAENASSL			68
DNA-binding protein H-NS family gi|24374659	38	14.6/5.56	2225.12	2226.16	PTVFKNEVNKGRSMoxDDFLI			58

[a] Name and NCBI Entrez accession number.
[b] Spot number according to the position on the 2D-PAGE.
[c] Mw calculated by using the residual monoisotopic values with cysteine converted to carbamidomethylcysteine.
[d] Mw observed in positive reflectron analysis (singly protonated).
[e] The sequence of the C-terminal peptide as found in the NCBI database, the part in bold was found by manual *de novo* sequencing.
[f] BLAST score obtained by searching the *de novo* determined sequence using standard settings *S. oneidensis* proteome (NCBI).
[g] MS homology score obtained by searching *S. oneidensis* proteins (NCBI) using the BLOSUM 62 scoring matrix.
[h] Protein score obtained from GPS explorer searching MS and MS/MS scores for protein identification. For proteins labeled M, scores were not significant but validated by manual interpretation at least one MS/MS spectrum.
[i] For proteins present in multiple spots, for simplicity, an examplary value for one of the spots is displayed.

Evaluation of the carboxypeptidase based method for C-terminal sequencing

To benchmark the automated method we also used our previously described manual technique, implementing CPase treatment [42], to analyze the 96 most intense spots from a 2D-PAGE gel containing *S. oneidensis* MR-1 proteins. Of the 96 spots, here 45 unique proteins could be identified by peptide mass fingerprint (PMF) analysis of the CNBr fragments and by performing a database search on a local MASCOT server using MS/MS data. In 16 spots, a sequence ladder could be obtained after incubation with CPase, corresponding to 8 different proteins. The length of the obtained sequence ladder varied between only 2 and 6 amino acids. Additional sequence information of the C-terminal peptide was generated by MS/MS. The obtained sequences were searched against the NCBI *S. oneidensis* MR-1 database using two different search engines (Table 3.3) significantly identifying 4 different proteins present in 13 spots.

To illustrate the workflow of the CPase technique we show the peptide map of cytoplasmic peptidyl-prolyl cis-trans isomerase B (spot 88) in detail. During the time-dependent ladder formation, samples were collected and spotted on the MALDI plate at different time points during the incubation with CPase. The peptide mass fingerprint analysis based on the spectrum from the sample before CPase incubation resulted in a significant protein identification score of 92 (result not shown). By combining MS spectra obtained from the concentration dependent digests a 4 amino acid C-terminal tag SVAA could be determined originating from the C-terminal peptide, with an m/z value of 1903.05 (Figure 3.10 A). The intact peptide was also selected for MS/MS. Now, a somewhat larger amino acid tag could be extracted from the spectrum, AVVIEKVS or SVKEIVVA (Figure 3.10 B). By combining this information with the C-terminal 4 amino acid tag SVAA the direction of the tag can be determined and combined to AVVIEKVSVAA. This sequence tag resulted in a positive identification in a BLAST search (Table 3.3).

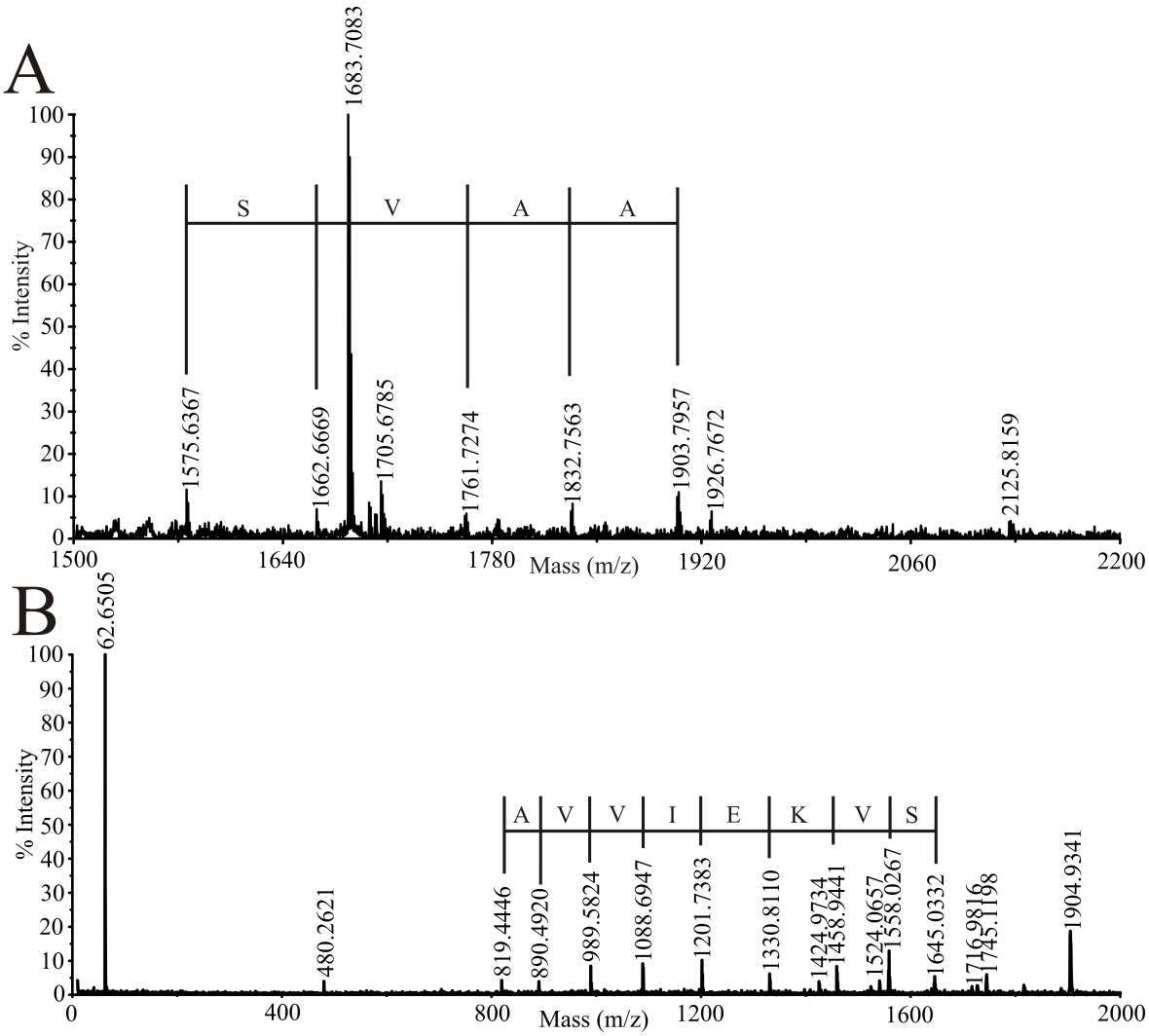


Figure 3.10: C-terminal sequence analysis of spot 88 (Cytoplasmic peptidyl-prolyl cistrans isomerase B PpiB) using carboxypeptidase (CPase) selection. Panel A: accumulated MS spectrum of the CNBr digest after incubation with CPase at different time points. A four amino acid long sequence ladder is formed starting at the C-terminal peptide (m/z 1903.79). Panel B: MS/MS spectrum of the C-terminal peptide. Only b-ions are observed.

Table 3.3: Carboxypeptidase based C-terminal sequence analysis of 2D PAGE-separated proteins of *Shewanella oneidensis*

Protein[a]	**Spot**[b]	**Mw (kDa) /pI**	**Mw calc. (Da)**[c]	**Mw obs. (Da)**[d]	**Sequence C-term peptide**[e]	**blastp score**[f]	**MS hom. score**[g]
Translation elongation factor Tu TufA gi|24371827	5-14	43.3/5.08	2554.43	2555.45	DEGLRFAIR**EGG**RTVGAGV-VAKIIA	1,3E-02	72
Translation elongation factor Tu TufB gi|24371815	5-14	43.3/5.13	2588.42	2589.40	DE**GLRFAIREGGRT**VGAG-VVAKIFA	1,0E-07	-
ABC-type tungstate uptake system substrate-binding component TupA gi|24376191	21	29.3/6.46	2088.80	2089.90	INS**FKVEGEQLFKATY**SE	2,0E-11	76
50S ribosomal protein L7/L12 RplL gi|24371821	23	12.5/4.63	3494.87	3495.30	SE**AAPVAVKEGVSKEEAE-ALKKELVEAG**ASVEIK	1,0E-20	121
Two component signal transduction system controlling aerobic respiration response regulator ArcA gi|24375475	53	27.2/5.51	5743.94	5744.19	TGRELKPHDRTVDVTIRRI-RKHFESLPDTPEIIATIHG-EGYRFCamGNLED	-	-
TonB2 energy transduction system periplasmic component gi|414561958	87	24.1/5.04	5200.78	5202.00	YNPQTKGWoxDKLEDSYLR-ELTKGIRIARKQGALDLFA-LPIPAAETAQ	5,8E-02	27
Cytoplasmic peptidyl-prolyl cis-trans isomerase B PpiB gi|24373356	88	18.1/5.30	1903.05	1904.1	HQDVPLE**AVVIEKV**SVAA	2,0E-06	46
Universal stress protein family gi|24375179	90	15.59/5.26	3665.95	3666.78	VVIASHGRTGISHFLHTNV-AEDVANGAVCPVLVVK	1,6E-01	21

[a] Name and NCBI Entrez accession number.
[b] Spot number according to the position on the 2D PAGE.
[c] Mw calculated by using the residual monoisotopic values with cysteine converted to carbamidomethylcysteine.
[d] Mw observed in positive reflectron analysis (singly protonated).
[e] The sequence of the C-terminal peptide as found in the NCBI database. The part in bold was found by manual *de novo* sequencing, the part in red was obtained via carboxypeptidase ladder formation.
[f] BLAST score obtained by searching the *de novo* determined sequence using standard settings *S. oneidensis* proteome (NCBI).
[g] MS homology score obtained by searching *S. oneidensis* proteins (NCBI) using the BLOSUM 62 scoring matrix.

3.3.5 Discussion

The chemical selection technique, automated on the robotic Tecan platform, outperforms our previous CPase based method in terms of throughput. 96 samples can be prepared for MALDI analysis starting from gel spots in 24 h (including a 12 h CNBr digest). The entire process only asks for two manual interventions. When samples are analyzed using the manual CPase technique a throughput of only 10 samples a day could be reached.

The MALDI data acquisition and interpretation are also faster and more straightforward when using the chemical selection technique. There is no need to analyze the samples at multiple time points or with multiple concentrations of CPase. This reduces the number of spots and spectra that need to be acquired and interpreted. The C-terminal peptide can be identified from a single MS spectrum while multiple MS spectra need to be layered to observe the ladder formed in the CPase technique.

When the results of both experiments are compared it is clear that the chemical selection technique has a larger number of identified and sequenced C-terminal peptides. There are two main reasons why the chemical selection would generate better results than the CPase selection. First, due to the selectivity of the CPase, a lot of C-terminal peptides are not efficiently digested and no ladder will be observed for those proteins, while the chemical selection is sequence independent. Second, when a sample is analyzed using the CPase approach the sample is distributed over multiple fractions (concentration- and time-dependent analysis) resulting in a lower concentration of the peptide of interest in the spot. Due to the ladder formation, the C-terminal peptide is also present in multiple forms, lowering its detection level.

We realize that 2D-PAGE has several well characterized limitations that limit the applicability of our technique, and that LC-MS based shotgun methods are in place [65]. However, 2D-PAGE is still the preferred method to separate intact proteins and offers the possibility to distinguish protein isoforms. Several proteins were identified in multiple spots, in all demonstrated cases providing knowledge that the isoforms are not the result of C-terminal processing.

The success of our C-terminal sequence determination depends on the length, ionization capacity and fragmentation efficiency of the CNBr fragments that are generated. In both cases the MS analysis is performed using MALDI-TOF MS, generating predominantly singly charged ions. The use of α-cyano-4-hydroxycinnamic acid as matrix produces chemical noise in the low molecular weight mass range and therefore hinders the detection of C-terminal peptides with a mass below 1 kDa. Further optimization of the MALDI sample preparation could probably improve this. It should be realized that performing C-terminal sequence analysis of tryptic endopeptides as obtained from typical shotgun methods theoretically results in over 70% of

peptides smaller than 1000 Da in *S. oneidensis*.

On the other hand, in our experience the upper mass limit for the analysis of CPase derived ladder sequences in MALDI-TOF MS, providing enough resolution and accuracy to identify the amino acid sequence, was restricted to C-terminal fragments with Mw of ±4-5 kDa. Similarly, MALDI-TOF MS/MS is typically restricted to peptides of the same length, limiting our new approach. Indeed, when analyzing the mass distribution of the C-terminal peptides generated by CNBr cleavage of all *S. oneidensis* MR-1 proteins, we calculated that 51.2% of the 4087 unique entries in the NCBI database have a C-terminal peptide mass between 1 kDa and 5.5 kDa and should be easily distinguished in a reflection MALDI-TOF MS spectrum. 70% of the proteins that we identified based on the peptide fingerprints, but for which no C-terminus could be obtained, have a predicted C-terminal peptide that falls outside this mass range. To be independent from genome annotation, the C-terminal sequence should be determined by *de novo* sequencing. The upper limit to generate qualitative *de novo* interpretable MS/MS spectra using our 4800 TOF/TOF instrument is around 3.5 kDa. Half of the C-termini that could not be identified by *de novo* sequence analysis, but were identified by a MASCOT MS/MS ion search (7 spots out of 14) were larger. It should be commented that MASCOT is not fully compatible with our strategy. For CNBr digested peptides, it automatically sets the C-terminal amino acid as variable for homoserine and homoserine lactone. It would be better to allow a fixed modification there, which would reduce the search window and result in more significant scores. This is also the reason why we report the GPS Scores, as this groups the scores for all fingerprint and MS/MS spectra of a particular MALDI spot analysis, providing a more confident identification.

Overall, we were unable to identify the C-terminal peptide in 23 identified spots (21 unique protein entries), around 70% of those C-termini were outside the 1-5.5 kDa range (Appendix Table A.1).

3.3.6 Conclusion

The determination of the actual C-termini of proteins on a proteome-wide scale is a challenging task. Common shotgun proteomic approaches often fail to characterize protein termini, especially C-termini. This is most often due to the poor detection of terminal peptides during mass spectrometric analysis of complex peptide mixtures and the incomplete annotation of protein termini in protein databases. Verification of predicted C-termini could be performed in LC-MS setups by generation of specific inclusion lists, based on the predicted mass of the C-terminal peptides. Here, we present the first fully automated C-terminal sequencing approach that can be implemented in a traditional proteomic setup. We have applied the method to 96 2D-PAGE separated *S. oneidensis* MR-1 proteins and show strong improvement compared to

our previously used manual CPase ladder sequencing technique. Moreover, we were able to identify three times more proteins using *de novo* sequenced C-terminal peptides. The main limitations of the technique are intrinsic to the use of 2D-PAGE and MALDI TOF/TOF MS as analysis tools. We have demonstrated that the technique theoretically covers 50% of the proteome of *S. oneidensis* MR-1 and is at least complementary to other approaches for whole genome C-terminal sequence determination.

3.3.7 Acknowledgements

B.S. is a Postdoctoral fellow of the Fund for Scientific Research-Flanders (F.W.O.-Vlaanderen, Belgium). P.M. and I.T. are funded by a Ph.D. grant of the institute for the promotion of Innovation through Science and Technology in Flanders (I.W.T.-Vlaanderen). The authors acknowledge funding by the Fund for Scientific Research-Flanders through Research Grant G.0644.07 (F.W.O.-Vlaanderen).

References

[1] Gross, E. and Witkop, B. (1962) Non-enzymatic cleavage of peptide bonds: the methionine residues in bovine pancreatic ribonuclease. *Journal of biological chemistry*, **237**, 1856–1860.

[2] Schreiber, J. and Witkop, B. (1964) The reaction of cyanogen bromide with mono- and diamino acids. *Journal of the American chemical society*, **86**, 2441–2445.

[3] Gross, E. and Witkop, B. (1961) Selective cleavage of the methionyl peptide bonds in ribonuclease with cyanogen bromide. *Journal of the American chemical society*, **83**, 1510–1511.

[4] Morrison, J., Fidge, N., and Grego, B. (1990) Studies on the formation, separation, and characterization of cyanogen bromide fragments of human AI apolipoprotein. *Analytical biochemistry*, **186**, 145–152.

[5] Kaiser, R. and Metzka, L. (1999) Enhancement of cyanogen bromide cleavage yields for methionyl-serine and methionyl-threonine peptide bonds. *Analytical biochemistry*, **266**, 1–8.

[6] Shively, J. E., Hawke, D., and Jones, B. N. (1982) Microsequence analysis of peptides and proteins: Artifacts and the effects of impurities on analysis. *Analytical biochemistry*, **120**, 312–322.

[7] Goodlett, D. R., Armstrong, F. B., Creech, R. J., and van Breemen, R. B. (1990) Formylated peptides from cyanogen bromide digests identified by fast atom bombardment mass spectrometry. *Analytical biochemistry*, **186**, 116–120.

[8] Loo, J. A., Quinn, J. P., Ryu, S. I., Henry, K. D., Senko, M. W., and McLafferty, F. W. (1992) High-resolution tandem mass-spectrometry of large biomolecules. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 286–289.

[9] Cordoba, O. L., Linskens, S. B., Dacci, E., and Santomé, J. A. (1997) 'In gel' cleavage with cyanogen bromide for protein internal sequencing. *Journal of biochemical and biophysical methods*, **35**, 1–10.

[10] Ambler, R. (1965) Behaviour of peptides formed by cyanogen bromide cleavage of proteins. *Biochemical journal*, **96**, P32.

[11] Knobler, Y., Bonni, E., and Sheradsky, T. (1964) Lactam formation through aminolysis of α-amino-γ-butyrolactone, 2-amino-4-hydroxybutyramides and 1-aryl 3-aminopyrrolidin-2-ones. *The journal of organic chemistry*, **29**, 1229–1236.

[12] Horn, M. J. and Laursen, R. A. (1973) Solid-phase Edman degradation: attachment of carboxyl-terminal homoserine peptides to an insoluble resin. *FEBS letters*, **36**, 285–288.

[13] Shi, T., Weerasekera, R., Yan, C., Reginold, W., Ball, H., Kislinger, T., and Schmitt-Ulms, G. (2009) Method for the affinity purification of covalently linked peptides following cyanogen bromide cleavage of proteins. *Analytical chemistry*, **81**, 9885–9895.

[14] Compagnini, A., Cunsolo, V., Foti, S., and Saletti, R. (2001) Improved accuracy in the matrix-assisted laser desorption/ionization-mass spectrometry determination of the molecular mass of cyanogen bromide fragments of proteins by post-cleavage reaction with tris (hydroxymethyl) aminomethane. *Proteomics*, **1**, 967–974.

[15] Horn, M. J. (1975) Amination of carboxyl-terminal homoserine peptides as an aid in peptide separation. *Analytical biochemistry*, **69**, 583–589.

[16] Murphy, C. M. and Fenselau, C. (1995) Recognition of the carboxy-terminal peptide in cyanogen-bromide digests of proteins. *Analytical chemistry*, **67**, 1644–1645.

[17] Yokoyama, S., Oobayashi, A., Tanabe, O., and Ichishima, E. (1975) Action of crystalline acid carboxypeptidase from *Penicillium janthinellum*. *Biochimica et biophysica acta - Enzymology*, **397**, 443–448.

[18] Hayashi, R. (1976) Carboxypeptidase Y. *Methods in enzymology*, **45**, 568–587.

[19] Jung, G., Ueno, H., and Hayashi, R. (1999) Carboxypeptidase Y: structural basis for protein sorting and catalytic triad. *Journal of biochemistry*, **126**, 1–6.

[20] Remington, S. J. (1993) Serine carboxypeptidases: a new and versatile family of enzymes. *Current opinion in biotechnology*, **4**, 462–468.

[21] Breddam, K. (1986) Serine carboxypeptidases. A review. *Carlsberg research communications*, **51**, 83–128.

[22] Bender, M. L., Schonbaum, G. R., and Zerner, B. (1962) Spectrophotometric investigations of the mechanism of α-chymotrypsin-catalyzed hydrolyses. detection of the acyl-enzyme intermediate. *Journal of the American chemical society*, **84**, 2540–2550.

[23] Fersht, A. (1985) *Enzyme structure and function*, vol. 10. WH Freeman & Co, New York.

[24] Stennicke, H. R., Mortensen, U. H., and Breddam, K. (1996) Studies on the hydrolytic properties of (serine) carboxypeptidase Y. *Biochemistry*, **35**, 7131–7141.

[25] Bech, L. M. and Breddam, K. (1989) Inactivation of carboxypeptidase Y by mutational removal of the putative essential histidyl residue. *Carlsberg research communications*, **54**, 165–171.

[26] Olesen, K., Meldal, M., and Breddam, K. (1996) Extended subsite characterization of carboxypeptidase Y using substrates based on intermolecularly quenched fluorescence. *Protein and peptide letters*, **3**, 67–74.

[27] Thiede, B., Wittmann-Liebold, B., Bienert, M., and Krause, E. (1995) MALDI-MS for C-terminal sequence determination of peptides and proteins degraded by carboxypeptidase Y and P. *FEBS letters*, **357**, 65–69.

[28] Myers, C. R. and Nealson, K. H. (1988) Bacterial manganese reduction and growth with manganese oxide as the sole electron acceptor. *Science*, **240**.

[29] Heidelberg, J. F., et al. (2002) Genome sequence of the dissimilatory metal ion–reducing bacterium *Shewanella oneidensis*. *Nature biotechnology*, **20**, 1118–1123.

[30] Daraselia, N., Dernovoy, D., Tian, Y., Borodovsky, M., Tatusov, R., and Tatusova, T. (2003) Reannotation of *Shewanella oneidensis* genome. *Omics: a journal of integrative biology*, **7**, 171–175.

[31] Romine, M. F., Elias, D. A., Monroe, M. E., Auberry, K., Fang, R., Fredrickson, J. K., Anderson, G. A., Smith, R. D., and Lipton, M. S. (2004) Validation of *Shewanella oneidensis* MR-1 small proteins by AMT tag-based proteome analysis. *Omics: a journal of integrative biology*, **8**, 239–254.

[32] Romine, M. F., Carlson, T. S., Norbeck, A. D., McCue, L. A., and Lipton, M. S. (2008) Identification of mobile elements and pseudogenes in the *Shewanella oneidensis* MR-1 genome. *Applied and environmental microbiology*, **74**, 3257–3265.

[33] Elias, D. A., Monroe, M. E., Marshall, M. J., Romine, M. F., Belieav, A. S., Fredrickson, J. K., Anderson, G. A., Smith, R. D., and Lipton, M. S. (2005) Global detection and characterization of hypothetical proteins in *Shewanella oneidensis* MR-1 using LC-MS based proteomics. *Proteomics*, **5**, 3120–3130.

[34] Elias, D. A., Monroe, M. E., Smith, R. D., Fredrickson, J. K., and Lipton, M. S. (2006) Confirmation of the expression of a large set of conserved hypothetical proteins in *Shewanella oneidensis* MR-1. *Journal of microbiological methods*, **66**, 223–233.

[35] Gupta, N., et al. (2007) Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome research*, **17**, 1362–1377.

[36] Deutschbauer, A., Price, M. N., Wetmore, K. M., Shao, W., Baumohl, J. K., Xu, Z., Nguyen, M., Tamse, R., Davis, R. W., and Arkin, A. P. (2011) Evidence-based annotation of gene function in *Shewanella oneidensis* MR-1 using genome-wide fitness profiling across 121 conditions. *Plos genetics*, **7**, e1002385.

[37] Wang, Z., Hilder, T. L., van der Drift, K., Sloan, J., and Wee, K. (2013) Structural characterization of recombinant α-1-antitrypsin expressed in a human cell line. *Analytical biochemistry*, **437**, 20 – 28.

[38] Jensen, O. N. (2006) Interpreting the protein language using proteomics. *Nature reviews molecular cell biology*, **7**, 391–403.

[39] Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G. R., and Vandekerck hove, J. (2003) Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nature biotechnology*, **21**, 566–569.

[40] McDonald, L., Robertson, D. H. L., Hurst, J. L., and Beynon, R. J. (2005) Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nature methods*, **2**, 955–957.

[41] Nakazawa, T., Yamaguchi, M., Okamura, T.-a., Ando, E., Nishimura, O., and Tsunasawa, S. (2008) Terminal proteomics: N- and C-terminal analyses for high-fidelity identification of proteins using MS. *Proteomics*, **8**, 673–685.

[42] Samyn, B., Sergeant, K., Castanheira, P., Faro, C., and Van Beeumen, J. (2005) A new method for C-terminal sequence analysis in the proteomic era. *Nature methods*, **2**, 193–200.

[43] Samyn, B., Sergeant, K., and Beeumen, J. V. (2006) A method for C-terminal sequence analysis in the proteomic era (proteins cleaved with cyanogen bromide). *Nature protocols*, **1**, 317–322.

[44] Patterson, D. H., Tarr, G. E., Regnier, F. E., and Martin, S. A. (1995) C-terminal ladder sequencing via matrix-assisted laser-desorption mass-spectrometry coupled with carboxypeptidase-Y time-dependent and concentration-dependent digestions. *Analytical chemistry*, **67**, 3971–3978.

[45] VerBerkmoes, N. C., Bundy, J. L., Hauser, L., Asano, K. G., Razumovskaya, J., Larimer, F., Hettich, R. L., and Stephenson, J. L. (2002) Integrating "top-down" and "bottom-up" mass spectrometric approaches for proteomic analysis of *Shewanella oneidensis*. *Journal of proteome research*, **1**, 239–252.

[46] Arthur, J. W. and Wilkins, M. R. (2004) Using proteomics to mine genome sequences. *Journal of proteome research*, **3**, 393–402.

[47] Armengaud, J. (2009) A perfect genome annotation is within reach with the proteomics and genomics alliance. *Current opinion in microbiology*, **12**, 292–300.

[48] Dormeyer, W., Mohammed, S., van Breukelen, B., Krijgsveld, J., and Heck, A. J. R. (2007) Targeted analysis of protein termini. *Journal of proteome research*, **6**, 4634–4645.

[49] Jornvall, H. (1977) Primary structure of yeast alcohol-dehydrogenase. *European journal of biochemistry*, **72**, 425–442.

[50] Meng, F., Cargile, B. J., Patrie, S. M., Johnson, J. R., McLoughlin, S. M., and Kelleher, N. L. (2002) Processing complex mixtures of intact proteins for direct analysis by mass spectrometry. *Analytical chemistry*, **74**, 2923–2929.

[51] Catherman, A. D., Durbin, K. R., Ahlf, D. R., Early, B. P., Fellers, R. T., Tran, J. C., Thomas, P. M., and Kelleher, N. L. (2013) Large-scale top-down proteomics of the human proteome: membrane proteins, mitochondria, and senescence. *Molecular & cellular proteomics*, **12**, 3465–3473.

[52] Kim, J.-S., Shin, M., Song, J.-S., An, S., and Kim, H.-J. (2011) C-terminal *de novo* sequencing of peptides using oxazolone-based derivatization with bromine signature. *Analytical biochemistry*, **419**, 211–216.

[53] Kosaka, T., Takazawa, T., and Nakamura, T. (2000) Identification and C-terminal characterization of proteins from two-dimensional polyacrylamide gels by a combination of isotopic labeling and nanoelectrospray Fourier transform ion cyclotron resonance mass spectrometry. *Analytical chemistry*, **72**, 1179–1185.

[54] Zhou, X. W., Blackman, M. J., Howell, S. A., and Carruthers, V. B. (2004) Proteomic analysis of cleavage events reveals a dynamic two-step mechanism for proteolysis of a key parasite adhesive complex. *Molecular & cellular proteomics*, **3**, 565–576.

[55] Sechi, S. and Chait, B. (2000) A method to define the carboxyl terminal of proteins. *Analytical chemistry*, **72**, 3374–3378.

[56] Xu, G., Shin, S. B. Y., and Jaffrey, S. R. (2011) Chemoenzymatic labeling of protein C-termini for positive selection of C-terminal peptides. *ACS Chemical Biology*, **6**, 1015–1020.

[57] Van Damme, P., Staes, A., Bronsoms, S., Helsens, K., Colaert, N., Timmerman, E., Aviles, F. X., Vandekerckhove, J., and Gevaert, K. (2010) Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nature methods*, **7**, 512–515.

[58] Schilling, O., Barre, O., Huesgen, P. F., and Overall, C. M. (2010) Proteome-wide analysis of protein carboxy termini: C terminomics. *Nature methods*, **7**, 508–U33.

[59] Chait, B. T., Wang, R., Beavis, R. C., and Kent, S. B. H. (1993) Protein ladder sequencing. *Science*, **262**, 89–92.

[60] Moerman, P., Sergeant, K., Debyser, G., Devreese, B., and Samyn, B. (2010) A new chemical approach to differentiate carboxy terminal peptide fragments in cyanogen bromide digests of proteins. *Journal of proteomics*, **73**, 1454–1460.

[61] Quadroni, M. and James, P. (1999) Proteomics and automation. *Electrophoresis*, **20**, 664–677.

[62] Carrette, O., Burkhard, P. R., Sanchez, J.-C., and Hochstrasser, D. F. (2006) State-of-the-art two-dimensional gel electrophoresis: a key tool of proteomics research. *Nature protocols*, **1**, 812–823.

[63] Perkins, D., Pappin, D., Creasy, D., and Cottrell, J. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

[64] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**, 3389–3402.

[65] Rabilloud, T., Chevallet, M., Luche, S., and Lelong, C. (2010) Two-dimensional gel electrophoresis in proteomics: past, present and future. *Journal of proteomics*, **73**, 2064–2077.

Chapter 4

Alternative cleavage reaction

4.1 Generating smaller C-terminal peptides

In the chemical selection strategy using CNBr, proteins are only cleaved C-terminal of methionine. Since methionine only accounts for 2.59% of the amino acids in *Shewanella oneidensis* MR-1, the peptides generated after CNBr cleavage are relatively large. In order for peptides to be detected on a MALDI-TOF/TOF analyzer they need to be in the 1-5.5 kDa mass range. Our data showed that most succesfully *de novo* interpreted MS/MS spectra had a parent ion with a molecular mass below 3.5 kDa. Using CNBr as cleavage reagent, 51% of the C-terminal peptides can be detected and only 1 out of 3 proteins generates a C-terminal peptide in the *de novo* sequencing mass range. To improve the proteome coverage of the technique an alternative cleavage reaction was optimized. When KI is added to the CNBr cleavage reaction mixture, next to cleavage C-terminal of methionine, also cleavage C-terminal to tryptophan is observed. Methionine is converted to a homoserine lactone (hsl), and tryptophan to a Cγ-O-spirolactone tryptophan [1]. The structural resemblance between the reaction products of the two amino acids, Cγ-spirolactone tryptophan being an indol-substituted homoserine lactone, suggested the likelihood of a similar behaviour during ladder sequencing by carboxypeptidases and chemical selection incubations. Tryptophan accounts for 1.25% of the amino acids in *S. oneidensis* MR-1. By cleaving C-terminal of tryptophan and methionine 58% of the C-terminal peptides are detectable on MALDI-TOF/TOF MS and 42% of the proteins have a C-terminal peptide in the *de novo* sequencing mass range.

4.2 Trp cleavage alternatives and reaction mechanisms

4.2.1 Different Trp cleavage methods

Oxidizing and halogenating agents are widely used for modification of amino acid side chains of proteins. [2–5] Some of these modifications were used for the non-enzymatic cleavage of

tryptophanyl peptide bonds in peptides and proteins. The oxidation of tryptophan during halogenation converts the indole group into oxindole and dioxindole in a combined oxidation and cleavage reaction. The first method widely used for the modification and cleavage of tryptophanyl peptide bonds utilized N-bromosuccinimide (NBS) as oxidant [6]. Because NBS is an extremely reactive agent, other amino acids are also modified under tryptophan cleavage conditions: tyrosine and histidine peptide bonds are cleaved during oxidation [7, 8], methionine is irreversibly converted to sulfone groups [9, 10] and cysteine and proline are oxidized [9]. In order to obtain a more selective cleavage, different strategies were used to create more mild reaction conditions. 8 M urea was added to form bromourea as a less reactive intermediate [11], and p-cresol and phenol were added to act as scavenger for tyrosine oxidation [12, 13]. Also milder brominating reagents were tested; (2-(2-nitrophenylsulfenyl)-3-methyl-3'-bromoindolenine (BNPS- skatole) [9], 2,4,6-tribromo-4-methylcyclohexadienone [14] and tribromocresol [15]. These reagents delivered a more selective cleavage of tryptophanyl bonds, but the other amino acids were still oxidized. Besides brominating reagents also other halonium releasing compounds were tested. Although iodine and chlorine cations are generally weaker oxidants than their bromine counterparts, they have sufficient potential to perform analogous cleavage reactions. The most important reagents used are N-iodosuccinimide (NIS) [16], N-chlorosuccinimide [10], N-chlorobenzotriazole [17], o-iodosobenzoic acid [13, 18], and chloramine-T combined with KI, I_2 and I_3^- [19]. N-chlorosuccinimide was reported to be the least harsh and only oxidized tryptophan and methionine [10, 20]. In 1977, Ozols *et al.* reported the combined cleavage of methionyl and tryptophanyl peptide bonds by performing CNBr cleavage in heptafluorobutyric acid [21].

4.2.2 Reaction mechanisms of Trp cleavage methods

Several slightly different reaction mechanisms for tryptophan oxidation and concomitant peptide bond cleavage after oxidative indole halogenation at low pH have been proposed over the years [6, 13, 14, 19]. The reaction products, stoichiometry and conversion speeds are monitored using UV-absorption at 280 nm during titration experiments. Patchornik *et al.* were the first to report the conversion of the indole spectrum in an oxindole spectrum similar to that of bromospiro-oxindole [6]. Later, the formation of an oxolactone product was confirmed by mass spectrometry [22, 23]. Slight differences were observed in stoichiometry between the different halogenation reagents used, all of them requiring 2 or 3 halogenating molecules per oxidation and cleavage reaction. The extra third halogenating molecule is used to form a stable halogen-oxindole derivative [14, 20]. In the reaction mechanism that was proposed by Patchornik, a bromonium intermediate is formed in the initial step. A nucleophilic attack of the carbonyl oxygen on the C3 carbon leads to the salt of an iminolactone, which after further oxidation gives a stable lactone (Figure 4.1) [6].

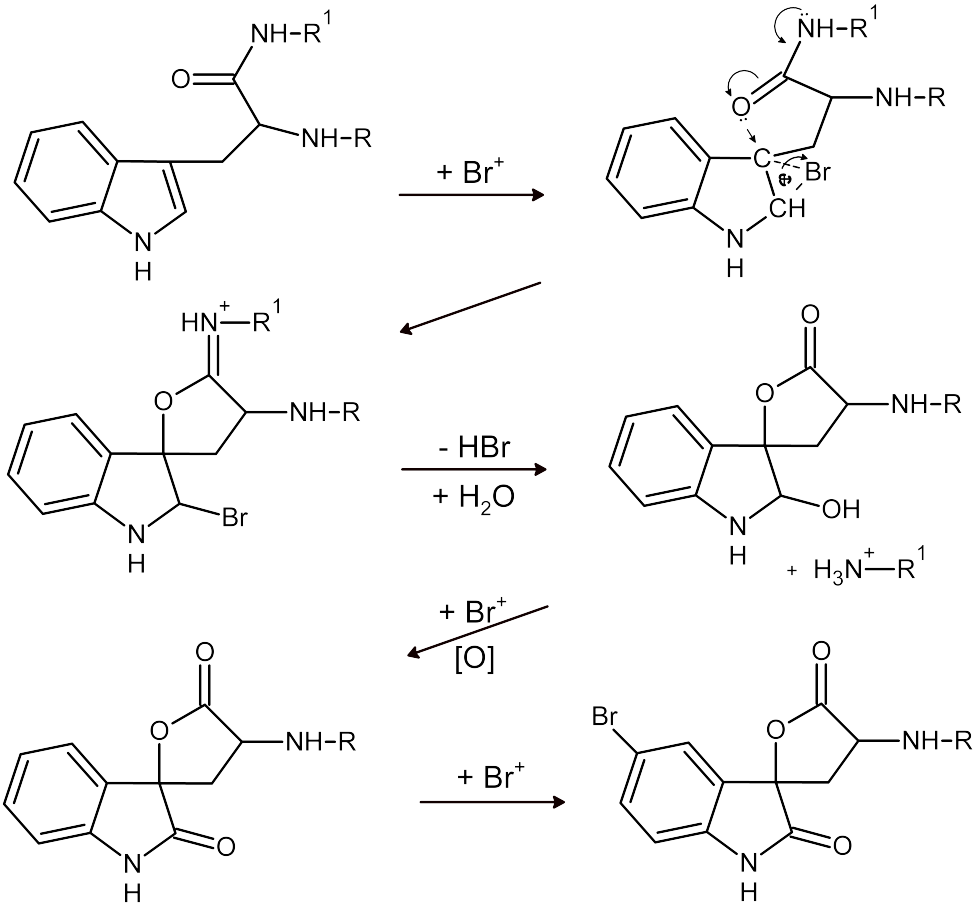


Figure 4.1: Reaction mechanism for oxidative halogenation of Trp according to Patchornik [14].

4.3 One-step chemical cleavage of tryptophanyl and methionyl peptide bonds with concomitant oxidation of disulfide bridges, usable in proteomic applications.

P.P. Moerman[1], K. Sergeant[2], B. Samyn[1], B. Devreese[1*]

[1] Laboratory for Protein Biochemistry and Biomolecular Engineering, Ghent University, Ghent, Belgium
[2] Department Environment and Agro-biotechnologies, Centre de Recherche Public, Gabriel Lippmann, Belvaux, Luxembourg

*Corresponding author: Bart.devreese@ugent.be

This chapter has been prepared for submission to Journal of Proteomics

4.3.1 Abstract

We present the optimized protocol to chemically cleave proteins C-terminal to methionine and tryptophan with concomitant cysteine oxidation using a CNBr and KI mixture. The characteristics, mechanism and reaction products of the cleavage were evaluated and the occurrence of possible side reaction products was studied with a panel of test peptides. Additionally, the spiro- and homoserine lactone groups were used to chemically discriminate between C-terminal and other peptides in the cleavage mixture. We were able to identify the C-terminal peptide in a digestion mixture of two test proteins and extend the application range of our previously reported C-terminal sequencing method.

4.3.2 Introduction

Contemporary proteomics is largely based on the separation of peptides or proteins followed by mass spectrometric analysis and database searches for protein identification. Although accurate mass determination and fragmentation of intact proteins can result in their identification [24, 25], most approaches depend on the application of specific and reproducible methods for cleavage of the protein into fragments of a size amenable to more routine mass spectrometric methods including peptide mass fingerprinting or/and tandem mass spectrometry based sequencing of peptides [26]. Proteins are therefore typically digested using endoproteases, most typically trypsin, but also endoproteinase LysC is applied. The rationale for their application is double. They leave a positively charged Lys or/and Arg C-terminal residue that is advantageous to generate more complete fragment ion series in MS/MS [27]. In addition, the abundance of these amino acid residues in proteins is moderate and, therefore, these enzymes generate peptide sizes that are suitable for mass spectrometric methods.

However, the latter is not the case for all proteins and for proteomic and other protein chemistry applications alternatives to these enzymes are still of interest. Enzymes with other cleavage specificities have proven quite useful, but in addition, the nature of certain amino acid residues has permitted the development of non-enzymatic, physical or chemical methods [28–31]. Although chemical cleavage methods form an alternative for endoprotease digestions, they are less frequently used, unless enzymes with the correct specificity are unavailable or when the use of enzymes is undesired for other reasons.

The most commonly used method for chemical cleavage, selective cleavage of methionyl peptide bonds with cyanogen bromide (CNBr), was first described in 1962 [32]. During the cleavage reaction the methionine is converted to homoserine in a pH-sensitive equilibrium with its lactone form (Figure 4.2 A), and the two forms can be quantitatively converted into each other [33]. Today this method, given the low abundance of methionine residues in proteins

typically resulting in larger peptides, is utilized to generate peptides for immunogenic studies [34]. Furthermore, CNBr cleavage is used to solubilize membrane proteins [35], and to digest proteins that are inaccessible to enzymes in studies of the disulfide bonding pattern of proteins [36, 37]. Proteins are typically cleaved with CNBr at a low pH and different reaction mixtures including 0.1 N HCl, 70% formic acid and 70% TFA have been used. Because formylation of side chains is avoided, the use of 70% TFA is preferred over the use of 70% formic acid [38], resulting in a nearly 100% cleavage yield. Met-Ser and Met-Thr bonds are cleaved less efficiently [39], but this impairment can be avoided by performing the cleavage in solutions containing a higher percentage of water, consequently a lower concentration of acid [40].

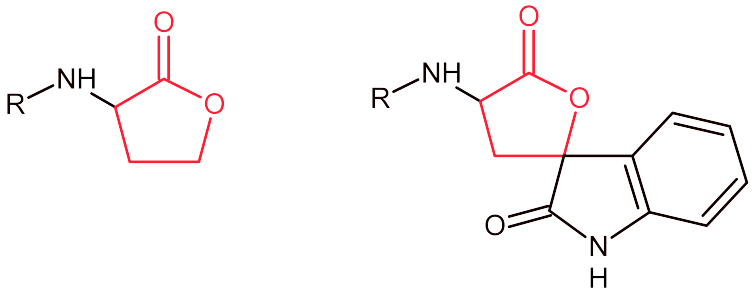


Figure 4.2: The different amino acid derivatives that are formed after digestion of a protein using KI/CNBr. A) After cleavage C-terminal to methionine the methionine is converted to a homoserine lactone in equilibrium with its open lactone form. B) Cleavage C-terminal to tryptophan converts this residue to an indol-substituted homoserine lactone, Cγ-O-spirolactone tryptophan.

In the previous chapters, we presented two C-terminal sequencing techniques using CNBr cleaved proteins. Carboxypeptidases and a chemical selection were applied to discriminate between the carboxylic acid ending C-terminal peptide and the peptides ending at a homoserine lactone. We also demonstrated that the methods can be used to study proteolytical processing events and can be performed on an automated robotic platform at a proteome-wide scale. Recently the technique has been used to determine the terminal amino acids of recombinant proteins [41].

During incubation with carboxypeptidases, only the C-terminal peptide is accessible to enzymatic degradation and forms a sequence ladder [42]. Alternatively, in the chemical approach, incubation of the CNBr mixture in a slightly basic buffer results in a partial opening of the homoserine lactone derivatives to the corresponding homoserine product (Δm = +18 Da). Therefore, all internal peptides appear as doublets in the MS spectrum, whereas the C-terminal fragment appears as the only singlet [43]. In both techniques, the C-terminal peptide can be distinguished and selected for MS/MS to unambiguously determine the C-terminal peptide. However, during our efforts to automate this methodology, we observed that many proteins

have C-terminal peptides with a molecular weight that is beyond routine mass limits to obtain good quality MS/MS data [44]. To reduce the size of C-terminal peptides, we were looking for an additional chemical method that could be introduced in our workflows.

Apart from chemical cleavage C-terminal of methionine, tryptophanyl peptide bonds can be cleaved by oxidative halogenation with 2-(2'-Nitrophenylsulphenyl)-3- methyl-3-bromoindoenine (BNPS-skatole), a nowadays rarely used chemical cleavage method [9]. Other reagents used for the cleavage of tryptophanyl peptide bond include N-bromosuccinimide, o-iodosobenzoic acid and dimethyl sulfide or aqueous HBr in acetic acid [2, 18, 45]. These methods result in the cleavage of the tryptophanyl peptide bond after oxidation of the tryptophan residue [13]. Contrary to the metabolic oxidation of tryptophan with reactive oxygen species [46], chemical oxidation of tryptophan with halogenated compounds results in the cleavage of peptide bonds, without opening the indole ring. Tryptophan is then oxidized to form Cγ-O-spirolactone tryptophan, a C4 oxidized indol substituted homoserine lactone (Figure 4.2 B) [6, 22]. Essentially the same reaction was also performed using an instrumental method: the electrochemical oxidation of proteins results in cleavage of polypeptides C-terminal to tyrosine and tryptophan [47].

In 1994, an abstract by Huang and Huang described a method for the concomitant chemical cleavage of tryptophanyl and methionyl peptide bonds using potassium iodide (KI) and CNBr. Proteins are cleaved at both sites and the fragments used for Edman degradation or separated by SDS-PAGE [1]. This method has been used a few times [48, 49]. It is thought that mixing KI and CNBr generates I_2, a reaction which has originally been proposed in a protocol for the iodometric determination of bromide in solutions [50, 51], and this consequently results in the formation of the spirolactone form of Trp and concomitant cleavage under acidic conditions [19]. Indeed, specific cleavage C-terminal of tryptophan was attained by incubating proteins in an excess of CNBr under oxidizing circumstances [21, 52]. Similar to what is observed in other oxidative halogenation reactions, we hypothesized that the method simultaneously results in the oxidation of cystine disulfides to two cysteic acid residues [14, 53].

Here, in view of applications in proteomic protocols, we optimized this procedure to cleave proteins C-terminal to both methionine and tryptophan, with concomitant cysteine oxidation. Using MS and MS/MS, the characteristics and mechanism of the cleavage were evaluated and the occurrence of possible side reaction products was studied with a set of test proteins. Additionaly, the spiro- and homoserine lactone groups were used to chemically discriminate between C-terminal and other peptides in the cleavage mixture. We were able to identify the C-terminal peptide in a digestion mixture of two test proteins and extend the application range of our previously reported chemical selection method [44].

4.3.3 Materials and methods

Materials

The four test proteins, used for optimization of the protocol, were mature avidin (*Gallus gallus*), cytochrome c (horse heart) and α-lactalbumin (*Bos taurus*) supplied by Sigma-Aldrich (Bornem, Belgium) and β-lactoglobulin (*Bos taurus*) supplied by Applied Biosystems (Framingham, MA, USA). Organic solvents, i.e. acetonitrile (ACN), methanol (MeOH) and ethanol, were from Biosolve (Valkenswaard, The Netherlands) and doubly deionized water was in-house purified with a Milli-Q water filtration system from Millipore (Bedford, MA, USA), further designated as MQ.

SDS-PAGE gel electrophoresis

The test proteins were run according to Laemmli in separate lanes on in-house casted 10 well 12% Tris-glycine gels (thickness 1 mm). Electrophoresis was carried out using a Mini-Protean 3 Cell (Bio-Rad, Nazareth, Belgium) at room temperature. 1/1 mixtures (v/v) of protein sample with sample buffer containing β-mercaptoethanol and bromophenol blue were heated briefly (95 °C, 5 min) and loaded on the gel. Electrophoresis, using 25 mM Tris base, 192 mM glycine, 0.1% SDS (w/v) as electrophoresis buffer, was carried out at 150 V until the bromophenol blue front reached the edge of the gel. The gels were fixated (2% H_3PO_4/50% ethanol; 30 minutes) and stained with Coomassie brilliant blue G-250 (0.2% in 34% MeOH/17% $(NH_4)_2SO_4$/3% phosphoric acid; 3 hours). The background of the gels was destained overnight in a 30% MeOH solution.

Destaining, reduction and alkylation

The separated proteins were excised and washed twice with 150 μl 200 mM NH_4HCO_3/50% ACN for 30 minutes at 30 °C and dried in a Speedvac (Thermo Savant, Holbrook, NY). The dried gel bands were submitted to reduction/alkylation, initially by incubation with 15 μl 10 mM dithiothreitol (DTT) in 7 M guanidinium HCl/0.3 M Tris, pH 9.0 (45 minutes at 55 °C). Alkylation was performed by adding 5 μl of a 10 mg/ml iodoacetamide and incubation in the dark for an additional 45 minutes at room temperature. For the reduction and alkylation of proteins in solution the same protocol was used, however the samples were desalted by using a ProSorb-device (Applied Biosystems) according to the manufacturer's instructions.

CNBr and CNBr/KI cleavage

CNBr-cleavage was performed according to a protocol previously described [42]. For CNBr cleavage in solution, dried protein was dissolved in 5 μl Milli-Q, 15 μl trifluoroacetic acid (TFA) (Applied Biosystems) and 5 μl 5 M CNBr in ACN (Sigma-Aldrich). After destaining and, when

applicable, reduction/alkylation, dried gel pieces were incubated with 5 μl MQ, 15 μl TFA and 5 μl 5 M CNBr in ACN. The samples were incubated overnight at 4 °C, after which proteins in solution were dried and reconstituted in 0.1% TFA in water. After incubation of gel-separated proteins, the supernatant was sequestered and peptides extracted from the gel by two washes with 35 μl 70% ACN/0.1% TFA, the solutions were pooled and dried. Prior to analysis the dried extracts were dissolved in 0.1% TFA.

For cleavage of methionyl and tryptophanyl peptide bonds, 4 μl MQ, 15 μl TFA (unless otherwise stated) and 5 μl 5 M CNBr in ACN were added to the dried protein or gel pieces. To this 1 μl of a 200 mM KI solution (unless otherwise stated) containing 4% phenol (Sigma-Aldrich) was added, reaching a final KI concentration of 8 mM. Samples were incubated overnight at 4 °C and further treated as described for CNBr-cleaved proteins.

To gain insight in the cleavage reaction an experiment was set up wherein horse heart cytochrome c was incubated with KI/CNBr in different ratios. In order to avoid the addition of excessive clean up steps, the concentration of CNBr was lowered and control digestions, using CNBr alone, were done in parallel. Incubations were performed with a KI/CNBr ratio of 2, 1, 0.5, 1/32, 1/63 and with the ratio used in the optimized protocol (1/125). Linear MALDI-TOF MS (see further) was used to analyze the results from these cleavages, for which the instrument was externally calibrated using the singly, doubly and triply charged ions from horse heart cytochrome c.

It must at all times be kept in mind that CNBr, TFA and phenol are toxic and corrosive products that can only be manipulated in a fume hood by skilled personnel, wearing protective clothing.

Incubations using other halogen salts were likewise performed, fluoro- chloro- and bromo-salts were added in the same ratio to CNBr as for the optimized KI-protocol described above, as were other iodo-salts.

Chemical opening

Before chemical opening of the lactone ring, the dried samples were desalted using C-18 micro purification tips (ZipTip, Millipore). The ZipTip protocol was performed as described by the supplier. A 50% ACN/0.1% TFA solution was used as activation and elution solvent. A 0.1% TFA solution in MQ was used as equilibration and washing solvent. After elution the samples were dried in a SpeedVac instrument. The partial opening of the homoserine- and Cγ-O-spirolactone was performed by dissolving the sample in 10 μl 12.5mM NH_4HCO_3 (pH 8.0), and incubation at 37 °C for 30 minutes.

Mass spectrometry

All mass spectra were acquired on an Applied Biosystems 4700 or 4800 TOF/TOF Proteomics analyzer. 0.7 μl aliquots of the samples were mixed with 0.4 μl of a 5 mg α-cyano-4-hydroxycinnamic/700 μl 50% ACN/0.1% TFA solution and applied on the MALDI plate, which was left to dry under ambient conditions.

For mass spectral analysis in reflectron mode, the mass spectrometer was externally calibrated with a mixture of angiotensin I, Glu-fibrinopeptide B, ACTH (18-39), ACTH (7-38) and des-Arg-bradykinin. For MS/MS experiments, the instrument was calibrated using Glu-fibrinopeptide fragments. MS/MS experiments were performed with the metastable suppressor on and 'gas off' in the fragmentation cell. Precursor ions were manually selected to ensure the fragmentation of those peptides of interest. For the fragmentation of peptides with a mass less than 3000 Da, the width of the mass window was set from -1 to +4 around the precursor peptide. For larger peptides this mass window was from -2 to +6.

Database searches were done using an in-house MASCOT platform. The cleavage-specificity C-terminal to tryptophan and methionine was programmed. For PMF-analysis, cleavage N-terminal to cysteic acid was neglected in the searches. Homoserine lactone and Cγ-O-spirolactone tryptophan were defined and set as variable modifications when C-terminal of peptides, cysteic acid was set as fixed modification.

4.3.4 Results

The aim of this study was to characterize in more detail the products obtained after a protocol that was proposed to perform chemical cleavage of proteins C-terminal to tryptophan and methionine simultaneously, and to implement it in our existing chemical selection technique for C-terminal sequencing [1, 43]. Similar to other oxidative halogenation reactions, we hypothesized that application of this protocol on proteins would also result in oxidation of cysteine and cystine to cysteic acid residues [14, 53]. From initial results it was clear that this was indeed the case, allowing the analysis of proteins after a single step, combining both cleavage of peptide bonds and complete oxidation of disulfides.

In the original protocol, a KI-solution is added to the reaction mixture simultaneously with the addition of CNBr. In order to avoid the missed cleavages, due to the inability of CNBr to react with oxidized methionine, it was tried to add the KI-solution to the sample after an initial incubation of 4 hours with CNBr alone. However, no impact of this initial incubation was apparent from the spectra (results not shown); this step was thus omitted in later experiments and KI was added simultaneously with CNBr. Formation of potassium-adducts was reduced by keeping the final concentration of KI to a minimum. Solutions from 500 mM to 25 mM were

tested and when 1 μl of a 200 mM solution was used, resulting in a final KI concentration of 8 mM, the cleavage remained optimal while the formation of K^+-adducts was minimized. Further lowering the concentration of KI resulted in missed cleavages at tryptophanyl peptide bonds. The protocol was further modified by adding trace amounts of phenol to the reaction mixture to avoid iodination of tyrosine. When 4% phenol was added to the halogen salt solution, the occurrence of this side reaction was completely eliminated.

The optimized protocol was then evaluated on four test proteins for which we used MALDI-TOF mass spectrometry (Figure 4.3). The major products are indeed peptides resulting from cleavage C-terminal to methionine and tryptophan, respectively ending on homoserine lactone (Δm= -48 Da) and Cγ-O-spirolactone tryptophan (Δm= +14 Da) (Figure 4.2, Table 4.1). When cysteine is present in the sequence, this is observed as cysteic acid (Δm= +48 Da). Some side reactions were observed, and especially oxidation of methionine or tryptophan without peptide bond cleavage resulted in missed cleavages. A more important side reaction was the cleavage of peptide bonds N-terminal to cysteine, when oxidized to cysteic acid. The actual chemical origin of this side reaction is unclear as it results in an additional loss of 1 Da for the N-terminal fragment.

Figure 4.3: MS-spectra of the test proteins incubated with CNBr/KI. Panel A: Horse heart cytochrome c. Panel B: β-lactoglobulin (*Bos taurus*). Panel C: Avidin (*Gallus gallus*). Panel D: α-lactalbumin (*Bos taurus*).

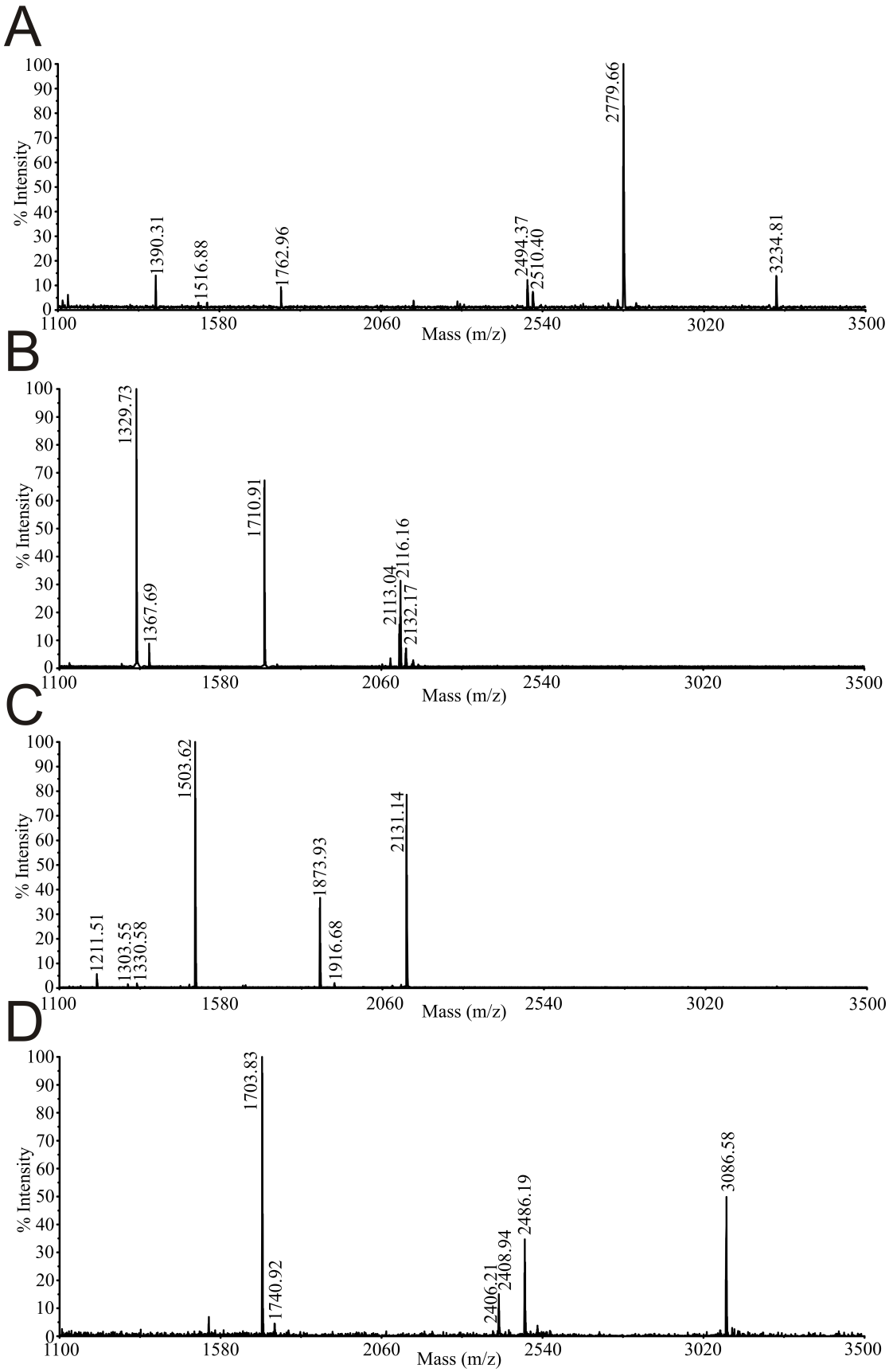

A
% Intensity
Mass (m/z)
1390.31
1516.88
1762.96
2494.37
2510.40
2779.66
3234.81
B
% Intensity
Mass (m/z)
1329.73
1367.69
1710.91
2113.04
2116.16
2132.17
C
% Intensity
Mass (m/z)
1211.51
1303.55
1330.58
1503.62
1873.93
1916.68
2131.14
D
% Intensity
Mass (m/z)
1703.83
1740.92
2406.21
2408.94
2486.19
3086.58

Table 4.1: Peptides observed after CNBr + KI digestion of cytochrome c (horse heart), β-lactoglobulin (*Bos taurus*), avidin (*Gallus gallus*) and α-lactalbumin (*Bos taurus*).

Mw calc. (Da)	**Mw obs. (Da)**	**Fragment**	**Sequence**	**Remarks**
Cytochrome c (horse heart)				
1390.28	1390.31	81 104	M.IFAGIKKKTEREDLIAYLKKATNE	MH_2^{++} of 2779.57
1517.87	1516.88	1 - 13	acGDVEKGKKIFVQK	X-Cys(14) cleavage
1762.93	1762.97	66 - 80	M.EYLENPKKYIPGTKM*	
2494.29	2494.37	60 - 80	W. KEETLMEYLENPKKYIPGTKM*	
2510.29	2510.40	60 - 80	W. KEETLM°EYLENPKKYIPGTKM*	
2779.57	2779.67	81 - 104	M.IFAGIKKKTEREDLIAYLKKATNE	C-term. peptide
3234.17	3234.31	1 - 59	acGDVEKGKKIFVQKCAQCHTVEKGGKH-KTGPNLHGLFGRKTGQAPGFTYTDANK-NKGITW*	N-term. $peptide^{2+}$
β-lactoglobulin (*Bos taurus*)				
1329.53	1329.73	8 - 19	M.KGLDIQKVAGTW*	
1367.54	1367.69	8 - 19	M.KGLDIQKVAGTW*	K^+ adduct of 1329.53
1711.89	1710.91	146-159	M.HIRLSFNPTQLEEQ	X-Cys(160) cleavage
2113.33	2113.04	146-162	M.HIRLSFNPTQLEEQCaHI	C-term. peptide
2116.52	2116.16	1-19	LIVTQTMKGLDIQKVAGTW*	N-term. peptide
2132.52	2132.17	1-19	LIVTQTM°KGLDIQKVAGTW*	N-term. peptide
Avidin (*Gallus gallus*)				
1211.38	1211.51	1-10	ARKCaSLTGKW*	N-term. peptide
	1303.55			n.i.
1331.44	1330.58	71 - 82	W.KFSESTTVFTGQ	X-Cys(83) cleavage
1503.60	1503.62	98 - 110	W.LLRSSVNDIGDDW*	
	1873.93			n.i.
	1916.68			n.i
2131.50	2131.14	111 - 128	W.KATRVGINIFTRLRTQKE	C-term. peptide
α-lactalbumin (*Bos taurus*)				
1703.84	1703.83	105 - 118	W.LAHKALCaSEKLDQW*	
1740.94	1740.92	91 - 104	M.CaVKKILDKVGINYW*	X-Cys(91) cleavage
	2408.94			n.i.
2486.22	2486.19	6 - 26	K.CaEVFRELKDLKGYGGVSLPEW*	X-Cys(6) cleavage
3086.55	3086.58	1 - 26	EQLTKCaEVFRELKDLKGYGGVSLPEW*	N-term. peptide

° Oxidation of Met of Trp.
* Homoserine lactone or Cγ-O-spirolactone tryptophan.
ac Acetylation.
Ca Cysteic acid.
n.i. not identified

Besides KI, other halogen salts and iodosobenzoic acid were evaluated as oxidative halogenation reagents. As presented in Table 4.2 for bovine β-lactoglobulin only the use of iodide salts resulted in the cleavage C-terminal to tryptophan and in the oxidation of cysteine. Although cleavage C-terminal to methionine was not affected, no cleavage at tryptophan was observed

when chloro- or bromo-salts were used, illustrating that I_2 formation is essential for the reaction. Although the peptide bond N-terminal to cysteic acid was sometimes cleaved, no random peptide bond fragmentation was observed under the conditions we applied (Table 4.2 & Supplementary Table B.1). Similar results were obtained for the three other proteins tested (Supplementary Table B.1). Because no differences were observed when using different iodide salts, the initial salt solution, 200 mM KI in 4% phenol/MQ, was used in experiments to establish whether using other reaction mixtures, previously used for CNBr-cleavage, could have an impact on the results. All reaction mixtures that were used; 70% TFA, 30% TFA, 30% formic acid, 0.1 N HCl and 0.05 N HCl, resulted in near identical spectra (results not shown). The incubation temperature and the incubation time seemed to have little effect on the results of the cleavage reaction. The results of these experiments for avidin are shown in Table 4.3.

Table 4.2: Peptides resulting from the cleavage of β-lactoglobulin using different halogen salts.

Fragment	Sequence	Remarks
KI		
1 - 19	LIVTQTM°KGLDIQKVAGTW	N-term. peptide
8 - 19	M.KGLDIQKVAGTW	
146 - 162	M.HIRLSFNPTQLEEQCaHI	C-term. peptide
146 - 159	M.HIRLSFNPTQLEEQ	X-Cys(160) cleavage
NaI		
1 - 19	LIVTQTM°KGLDIQKVAGTW	N-term.peptide
8 - 19	M.KGLDIQKVAGTW	
146 - 162	M.HIRLSFNPTQLEEQCaHI	C-term. peptide
146 - 159	M.HIRLSFNPTQLEEQ	X-Cys(160) cleavage
4-Iodosobenzoic acid		
1 - 19	LIVTQTM°KGLDIQKVAGTW	N-term. peptide
8 - 19	M.KGLDIQKVAGTW	
146 - 162	M.HIRLSFNPTQLEEQCaHI	C-term. peptide
KCl/NaCl/4-chlorobenzoic acid		
8 - 24	M.KGLDIQKVAGTWYSLAM	
146 - 162	M.HIRLSFNPTQLEEQCHI	C-term. peptide
KBr/NaBr/2-bromobenzoic acid		
1 - 24	M.LIVTQTMKGLDIQKVAGTWYSLAM	Trp or Met oxidation
8 - 24	M.KGLDIQKVAGTW°YSLAM	
146 - 162	M.HIRLSFNPTQLEEQCHI	C-term. peptide

Terminal Met or Trp residues were observed in the homoserine- or Cγ-O-spirolactone form.
°= Oxidation of Met of Trp.
Ca = Cysteic acid.

Table 4.3: Peptides observed after cleavage of avidine (*Gallus gallus*) at different temperatures and after incubation at 4 °C from 4 to 24 hours.

Mw Obs.	Sequence	Mw Obs.	Sequence
Overnight at -20 °C		**4 hours at 4 °C**	
1211.63	ARKCaSLTGKW	1211.61	ARKCaSLTGKW
1503.78	M.W°LLRSSVNDIGDDW	2966.53	W.KFSESTTVFTGQCaFIDRNGKEVLKTM
1721.85	W.LLRSSVNDIGDDW	3230.60	W.KFSESTTVFTGQCaFIDRNGKEVLKTM°W
2131.36	W.KATRVGINIFTRLRTQKE	1721.85	M.W°LLRSSVNDIGDDW
		1503.71	W.LLRSSVNDIGDDW
Overnight at 4 °C		2131.29	W.KATRVGINIFTRLRTQKE
1211.63	ARKCaSLTGKW		
2966.62	W.KFSESTTVFTGQCaFIDRNGKEVLKTM	**8 hours at 4 °C**	
3230.69	W.KFSESTTVFTGQCaFIDRNGKEVLKTM°W	1211.61	ARKCaSLTGKW
1503.78	M.W°LLRSSVNDIGDDW	2966.53	W.KFSESTTVFTGQCaFIDRNGKEVLKTM
1721.85	W.LLRSSVNDIGDDW	3230.60	W.KFSESTTVFTGQCaFIDRNGKEVLKTM°W
2131.36	W.KATRVGINIFTRLRTQKE	1721.85	M.W°LLRSSVNDIGDDW
		1503.71	W.LLRSSVNDIGDDW
Overnight at 24 °C		2131.29	W.KATRVGINIFTRLRTQKE
1211.63	ARKCaSLTGKW		
3230.69	W.KFSESTTVFTGQCaFIDRNGKEVLKTM°W	**16 hours at 4 °C**	
1503.78	M.W°LLRSSVNDIGDDW	1211.61	ARKCaSLTGKW
1721.85	W.LLRSSVNDIGDDW	2966.53	W.KFSESTTVFTGQCaFIDRNGKEVLKTM
2131.36	W.KATRVGINIFTRLRTQKE	3230.60	W.KFSESTTVFTGQCaFIDRNGKEVLKTM°W
		1721.85	M.W°LLRSSVNDIGDDW
		1503.71	W.LLRSSVNDIGDDW
		2131.29	W.KATRVGINIFTRLRTQKE
		20 hours at 4 °C	
		1211.61	ARKCaSLTGKW
		1503.71	W.LLRSSVNDIGDDW
		2131.29	W.KATRVGINIFTRLRTQKE
		24 hours at 4 °C	
		1211.61	ARKCaSLTGKW
		1503.71	W.LLRSSVNDIGDDW
		2131.29	W.KATRVGINIFTRLRTQKE

Terminal Met or Trp residues were observed in the modified homoserine lactone or Cγ-O-spirolactone form.
° Oxidation of Met of Trp.
Ca Cysteic acid.

Finally, the effects of incubation using different molar ratios of CNBr/KI were studied. The results of these experiments for horse heart cytochrome c are discussed below in detail and are represented in Figure 4.4 and Table 4.4. A first observation is that there is a 55 Da mass shift for the intact protein between samples incubated with and without KI. For the apocytochrome c this mass shift is lost (Table 4.4, experiment A1 and B1; Figure 4.4, panel B and C). This suggests that Fe is displaced from the heme-group when KI is added to the reaction mixture. Starting from the lowest concentration of CNBr, cleavages C-terminal to methionine can be observed, as is illustrated by the presence of a peak at m/z 4572 corresponding to peptide 66-104 in experiment B1 (Figure 4.4, panel B). In experiments B3-B6, where the final concentration CNBr surpassed 16 mM, the spectra are similar and no extra peaks can be observed. No cleavages were observed when proteins were incubated in 60% TFA/8 mM KI alone. However, the peptide bond

C-terminal to the only tryptophan present in the protein is cleaved at equimolarity between KI and CNBr (Figure 4.4, panel D). A first peak of very low intensity corresponding to cleavage C-terminal to methionine is only observed when a twofold excess of CNBr is added (Result only visible in reflector mode and not shown). At higher concentrations of CNBr, all possible cleavage sites are attached and even a 125-fold excess of CNBr does not seem to impair the cleavage C-terminal to tryptophan (Figure 4.4, Panel E). When the concentration of CNBr increases, so is the oxidation of methionine when present internally in a peptide sequence [49]. This is seen for the peptide KEETLMEYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE (sequence 60-104; calculated average m/z 5305) observed in panel D at m/z 5304. In panel E this peptide is dominantly present containing a single oxidized methionine and two oxidized methionines, respectively at m/z 5320 and 5336. Similar results were obtained for the other three test proteins (results not shown).

Figure 4.4: Linear mode MS spectra (4-15 kDa) after incubation of native horse heart cytochrome C with different ratios of KI/CNBr. Panel A: no KI/CNBr. Panel B: only CNBr, no KI (Table 4.4 experiment B3). Panel C: KI/CNBr 2/1 (Table 4.4 experiment A1). Panel D: equimolarity (Table 4.4 experiment A2). Panel E: KI/CNBr = 1/125 (Table 4.4 experiment A6). The peaks in the spectra are labeled with their m/z, more information on resulting peptides.

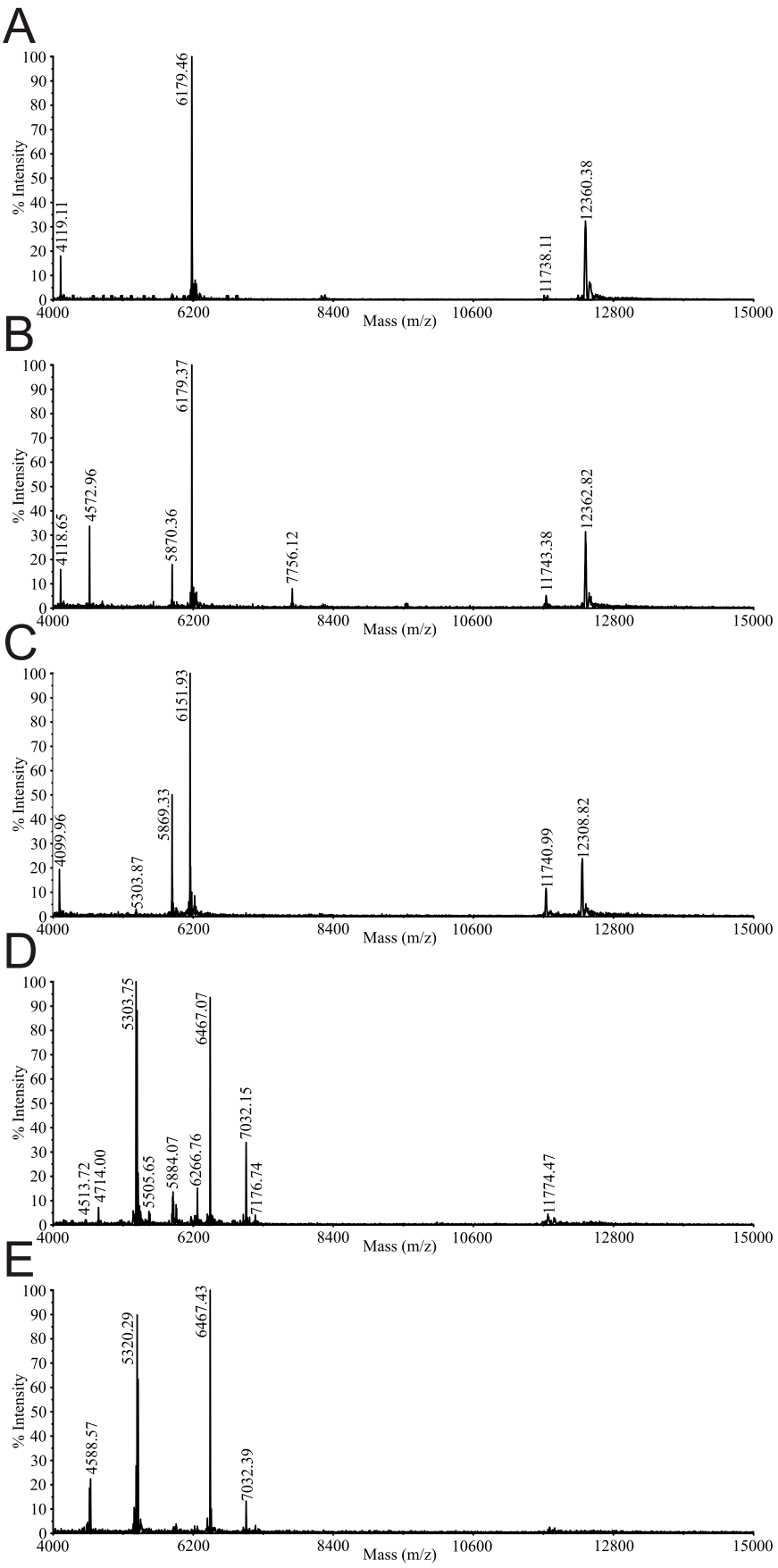

A
% Intensity
Mass (m/z)
4119.11
6179.46
11738.11
12360.38
B
% Intensity
Mass (m/z)
4118.65
4572.96
5870.36
6179.37
7756.12
11743.38
12362.82
C
% Intensity
Mass (m/z)
4099.96
5303.87
5869.33
6151.93
11740.99
12308.82
D
% Intensity
Mass (m/z)
4513.72
4714.00
5303.75
5505.65
5884.07
6266.76
6467.07
7032.15
7176.74
11774.47
E
% Intensity
Mass (m/z)
4588.57
5320.29
6467.43
7032.39

Table 4.4: Peptides observed after incubation of native horse heart cytochrome c with different ratio of KI/CNBr.

Mw Obs. (Da)	Fragment	Cleavage site	Mw Obs. (Da)	Fragment	Cleavage site
A1: 4 mM CNBr/8 mM KI			**B1: 4 mM CNBr**		
12310	holocyt-Fe	/	12365	holocyt	/
11743	apocyt	/	11744	apocyt	/
5304	60-104	W	4572	66-104	M
			7754	1-65	M
A2: 8 mM CNBr/8 mM KI			**B2: 8 mM CNBr**		
7032	1-59	W	12363	holocyt	/
6467	1-59	W	11744	apocyt	/
5304	60-104	W	4573	66-104	M
			7754	1-65	M
A3: 16 mM CNBr/8 mM KI			**B3: 16 mM CNBr**		
7032	1-59	W	7754	1-65	M
6467	1-59	W	7136	1-65	M
5304	60-104	W	4573	66-104	M
2779,83	81-104	M	2779,83	81-104	M
			1763,05	66-80	M
A4: 250 mM CNBr/8 mM KI			**B4: 250 mM CNBr**		
7032	1-59	W	7754	1-65	M
6467	1-59	W	7136	1-65	M
5320	60-104	W	4573	66-104	M
4588	66-104	M	2779,83	81-104	M
2779,83	81-104	M	1763,05	66-80	M
2494,52	60-80	M			
A5: 500 mM CNBr/8 mM KI			**B5: 500 mM CNBr**		
7032	1-59	W	7754	1-65	M
6467	1-59	W	7136	1-65	M
5320	60-104	W	4573	66-104	M
4588	66-104	M	2779,83	81-104	M
2779,83	81-104	M	1763,05	66-80	M
2494,52	60-80	M			
1763,05	66-80	M			
A6: 1000 mM CNBr/8 mM KI			**B6: 1000 mM CNBr**		
7032	1-59	W	7754	1-65	M
6467	1-59	W	7136	1-65	M
5320	60-104	W	4573	66-104	M
4588	66-104	M	2779,83	81-104	M
2779,83	81-104	M	1763,05	66-80	M
2494,52	60-80	M			

Masses below 4000 Da were observed in reflector mode, masses above 4000 Da were observed in linear mode.

In order to identify the C-terminal peptide in the CNBr/KI digest, both the homoserine lactone and γ-spirolactone ending peptides have to be present in their lactonised and hydrolysed form simultaneously. All internal peptides will appear as doublets (Δ mass = 18 Da) in the MALDI MS spectrum, while the carboxyl group ending C-terminal peptides will appear as a singlet. Different reaction conditions were tested on two test proteins that should produce a C-terminal peptide with a molecular mass in the range of MALDI-TOF/TOF MS analysis; β-lactoglobulin (*Bos taurus*) and avidin (*Gallus gallus*). Similar to the previously reported lactone ring opening protocol, 12.5 mM NH_4HCO_3, pH 8, was used as slightly basic buffer solution [43]. The digested samples were incubated at RT and 37 °C for 30 minutes, 1 hour and 24 hours. The best results were obtained at the elevated temperature and after minimum 1 hour. Longer incubation times resulted in a higher percentage of K^+-adduct formation (+38 Da). After cleavage and incubation with the basic buffer of the β-lactoglobulin and avidin fragments, the C-terminal peptides (2112.86 and 2131.13 Da) and multiple internal peptides were observed in the MS spectra (Figure 4.5, 4.6 and Table 4.5, 4.6). The only missed cleavages observed were due to oxidized Met and Trp residues. Hsl and spirolactone ending peptides were observed in the spectrum in the open and ring form simultaniously. The C-terminal peptide of the avidin fragment, observed as singlet in the spectra was selected for MS/MS and *de novo* sequenced (Figure 4.6). Important to note is that the C-terminus is not the only singlet present in the spectrum. Due to the partial cleavage N-terminally of cysteic acid, a second peptide is formed that does not end at a lactone ring structure and is observed as a singlet (1710.75 and 1330.59). Since this side reaction is not quantitative, the intact peptide was also visible in the spectrum (2112.86 and 2966.35). MS/MS interpretation allowed to distinguish this as the false positive from the real C-terminal peptide.

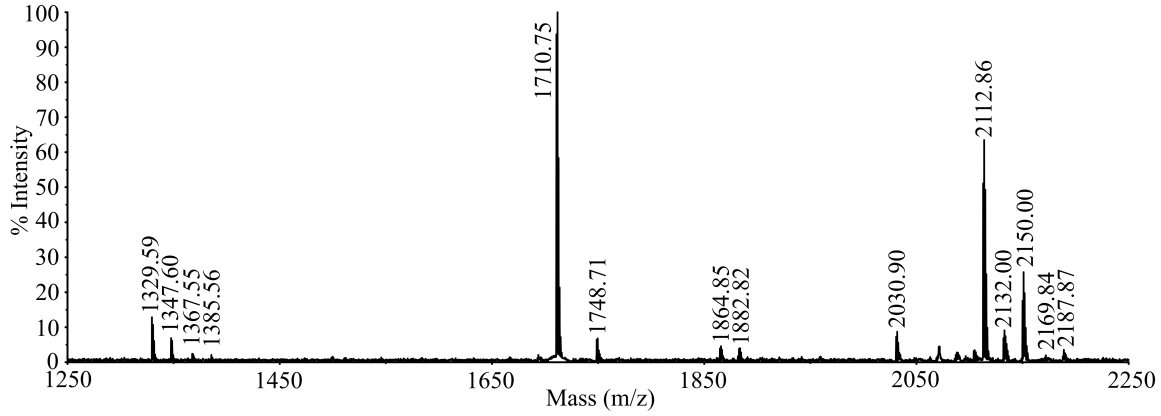


Figure 4.5: MALDI MS analysis of CNBr/KI digested β-lactoglobulin (*Bos taurus*) after partial homoserine lactone and γ-spirolactone ring opening. The C-terminal peptide is observed at 2112.87 Da. All peaks in the spectrum are annotated in Table 4.5.

Table 4.5: Peptides of β-lactoglobulin (*Bos taurus*) present after homoserinelactone and spirolactone ringopening

Mw calc. (Da)	**Mw obs. (Da)**	**Fragment**	**Sequence**	**Remarks**
1329.53	1329.60	8 - 19	M.KGLDIQKVAGTW*	
1347.54	1347.61	8 - 19	M.KGLDIQKVAGTW**	
1367.62	1367.55	8 - 19	M.KGLDIQKVAGTW*	K^+ adduct of 1329.53
1385.63	1385.56	8 - 19	M.KGLDIQKVAGTW**	K^+ adduct of 1347.54
1711.89	1710.76	146-159	M.HIRLSFNPTQLEEQ	X-Cys(160) cleavage
1749.98	1748.72	146-159	M.HIRLSFNPTQLEEQ	K^+ adduct of 1711.89
1865.22	1864.86	8 - 24	M.KGLDIQKVAGTW°YSLAM*	
1883.23	1882.83	8 - 24	M.KGLDIQKVAGTW°YSLAM**	
	2030.90			n.i.
2113.33	2112.87	146 - 162	M.HIRLSFNPTQLEEQCaHI	C-term. peptide
2132.52	2132.00	1 - 19	LIVTQTM°KGLDIQKVAGTW*	
2150.53	2150.00	1 - 19	LIVTQTM°KGLDIQKVAGTW**	
2170.61	2169.84	1 - 19	LIVTQTM°KGLDIQKVAGTW*	K^+ adduct of 2132.52
2188.62	2187.97	1 - 19	LIVTQTM°KGLDIQKVAGTW**	K^+ adduct of 2150.53

° Oxidation of Met of Trp.
* Homoserine lactone or Cγ-O-spirolactone tryptophan.
** Homoserine or opened Cγ-O-spirolactone tryptophan.
Ca Cysteic acid.
n.i. not identified

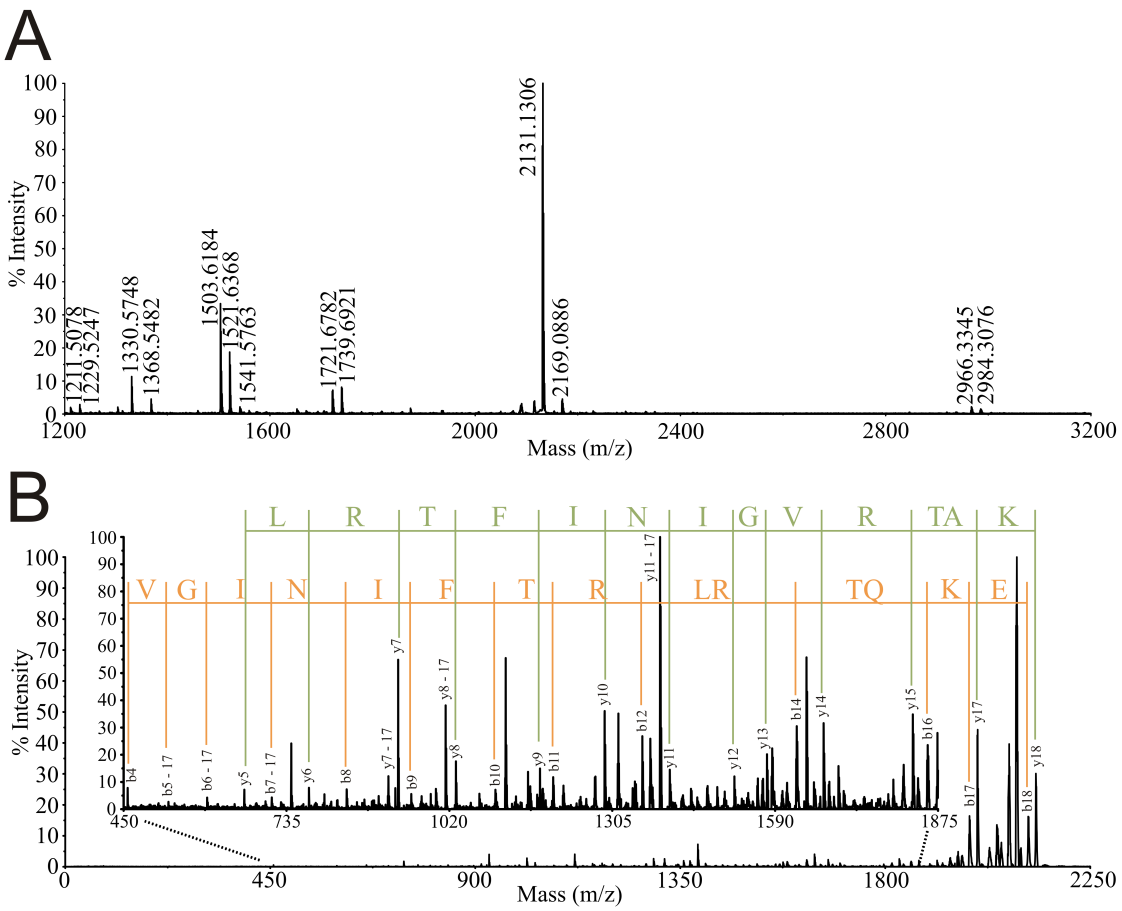


Figure 4.6: C-terminal sequence analysis of avidin (*Gallus gallus*) using partial homoserine lactone and γ-spirolactone ringopening after CNBr/KI digest to select the C-terminal peptide. Panel A: MALDI MS spectrum of the CNBr/KI generated avidin peptides. The C-terminal peptide is observed at 2131.13 Da and is the only singlet in the spectrum. All internal peptides are visible as doublets (Δ mass = 18 Da). Most peptides are also detected as K^+ adduct (Δ mass = 38 Da). All peaks in the spectrum are annotated in Table 4.6. Panel B: MALDI MS/MS spectrum of the C-terminal peptide at m/z 2131.13. Y- and b-ions are indicated in green and red respectively. The amino acid sequence is indicated in the one-letter code.

Table 4.6: Peptides of avidin (*Gallus gallus*) present after homoserinelactone and spirolactone ringopening

Mw calc. (Da)	Mw obs. (Da)	Fragment	Sequence	Remarks
1211.38	1211.51	1 10	ARKCaSLTGKW*	
1229.39	1229.52	1 10	ARKCaSLTGKW**	
1331.44	1330.58	71 - 82	W.KFSESTTVFTGQ	X-Cys(83) cleavage
1369.53	1368.55	71 - 82	W.KFSESTTVFTGQ	K^+ adduct of 1330.59
1503.60	1503.62	98 - 110	W.LLRSSVNDIGDDW*	
1521.61	1521.64	98 - 110	W.LLRSSVNDIGDDW**	
1541.70	1541.58	98 - 110	W.LLRSSVNDIGDDW*	K^+ adduct of 1503.62
1559.71	1559.62	98 - 110	W.LLRSSVNDIGDDW**	K^+ adduct of 1521.64
1721.81	1721.68	97 - 110	M.W°LLRSSVNDIGDDW*	
1739.82	1739.70	97 110	M.W°LLRSSVNDIGDDW**	
2131.50	2131.14	111 - 128	W.KATRVGINIFTRLRTQKE	C-term. peptide
2169.59	2169.09	111 - 128	W.KATRVGINIFTRLRTQKE	K^+ adduct of 2131.14
2967.40	2966.32	71 - 96	W.KFSESTTVFTGQCaFIDRNGKEVLKTM*	
2985.41	2984.31	71 96	W.KFSESTTVFTGQCaFIDRNGKEVLKTM**	

° Oxidation of Met of Trp.
* Homoserine lactone or Cγ-O-spirolactone tryptophan.
** Homoserine or opened Cγ-O-spirolactone tryptophan.
Ca Cysteic acid.

4.3.5 Discussion

During the application of CNBr to generate peptide fragments, it became apparent that the low occurrence of methionine residues limits the use of the previously described protocols for C-terminal sequencing [42, 44]. The need to include specific manipulations to ensure opening of disulfide bonds, especially for proteins in solution or after separation with SDS-PAGE, further complicates the protocol. These limitations prompted us to search for alternative digestion methods with which the conversion of the C-terminal amino acid of internal peptides, to a group that is not cleaved off by incubation with carboxypeptidases, is maintained. The chemical cleavage C-terminally of two different residues simultaneously, namely methionine and tryptophan, made the use of the protocol described by Huang and Huang an attractive alternative [1]. Furthermore, the structural resemblance between the reaction products of the two amino acids, Cγ-spirolactone tryptophan being an indol-substituted homoserine lactone, suggested of a similar behaviour during carboxypeptidase incubations and ring opening experiments. The oxidative environment required for cleavage C-terminally of tryptophan was thought to be sufficiently strong to also allow the oxidation of disulfide bonds, thereby eliminating the need for their reduction and alkylation.

A known side reaction, halogenation of tyrosine (when using KI; Δm= +125.9 Da) [54], was avoided by adding phenol to the reaction mixture. Phenol has been previously used as a

scavenger for free iodide [12]), and the addition of 4% phenol to the salt solution (final concentration in the reaction mixture 0.16%) completely eliminated this side reaction. An optimal concentration of KI was determined.

As can be seen in Figure 4.3, the number and the impact of modifications of amino acid side chains are limited. The conditions during the KI-cleavage are strongly oxidative, resulting in the complete oxidation of cysteine to cysteic acid when the incubation is done without previous alkylation of the sulfhydryl-moiety (Table 4.1). Nevertheless, aside from the oxidation of cysteine little or no oxidative damage to amino acid side chains was observed. Some of the side reactions are probably difficult to avoid and have been observed in most studies using chemical means to cleave peptide bonds C-terminally of tryptophan. The side reaction that is most troublesome is the peptide bond cleavage N-terminally of cysteic acid and the concomitant formation of uncharacterized reaction products. Although this can be avoided by protecting the sulfhydryl-moiety prior to the chemical cleavage, the benefit of simultaneous peptide bond cleavage and disulfide oxidation is then lost. No reference to this type of peptide bond cleavage is known to our knowledge. To assess the possible negative impact that the oxidation of cysteine and the side reactions have on the general applicability of the cleavage method described, more extensive sets of data need to be obtained. Nonetheless, the described cleavage method can be applied for general proteomic applications offering a single-incubation protocol to attain cleavage and the opening of disulfide bridges.

Unfortunately, the implementation of the technique in standard proteomic setups is slightly hindered by the limitations of the protein identification algorithms such as MASCOT. Currently, MASCOT does not allow to define a fixed modification at the cleavage site. Due to the chemical mechanism of the cleavage reaction, methionine and tryptophan are converted to respectively homoserine lactone and Cγ-O-spirolactone tryptophan. These modifications can only be selected as variable modifications. This has several consequences: false positive hits are created since peptides ending at an unmodified Met and Trp are allowed, multiple modifications reduce the score of a peptide match, and multiple variable modifications increase search times. Furthermore, the unwanted cleavage N-terminally of cysteine results in non-sense protein identifications. Similarly, missed cleavages are mostly induced by oxidized Met and Trp residues. Including this knowledge to restrict the MASCOT searching would improve the quality of the protein identifications. To circumvent the current limitations, all Met and Trp residues in the protein library can be altered into user defined amino acids that account for the generation of these derivatives. By analyzing the samples in an acidic environment only the closed lactone form should be present in the spectrum and identification would then be more performant.

When the KI-solution is added to the reaction mixture a brown-reddish color is formed. The reaction mechanism therefore was hypothesized to commence with the reaction of CNBr with I^- leading to the formation of active iodine species, I_2, and I^+, that subsequently donate iodide to susceptible groups. The mechanism of the tryptophan oxidation and peptide bond cleavage by active iodine species is proposed to follow a similar reaction mechanism to that proposed by Patchornik in 1960 for cleavage of tryptophan by N-bromosuccinimide [6, 19]. It is obvious that both KI and CNBr must be present to cleave peptide bonds C-terminal to tryptophan (Table 4.4, exp B). Furthermore, a molar excess of CNBr must exist before initiating the cleavage at both methionine and tryptophan residues wheras peptide bonds C- terminally of tryptophan alone are cleaved from the moment KI is added in the presence of stoichiometric amounts of CNBr. These reactions are independent of the solvents used to create an acidic environment, and of the incubation temperature and of time (Table 4.3). Although we used almost exclusively KI, identical results were obtained using other iodide-salts, including cleavage at both residues and oxidation of cysteine-bridges. Adding identical concentrations of other halogen salts, namely bromide- and chloride-salts, do not impair the cleavage at methionyl peptide bonds, but do not result in cleavage after Trp (Table 4.2).

Although iodide seems to be the central element in the reactions observed, fairly little has been published that links iodide to disulfide oxidation [14] and the concomitant cleavage N-terminally of cysteic acid. However, it has been noted that iodination of tyrosine residues with NaI, using chloramine T, results in the oxidation of cysteine to cysteic acid [55]. A similar oxidation was noticed when performing CNBr-cleavages of cytochrome c after alkylation of the cysteines with iodoacetamide [56]. Lederer and Tarin further developed the observed oxidation of cysteine to a method for the cleavage of the heme-group from cytochrome c [56], a reaction we also observed both with and without the addition of KI to the CNBr reaction mixture (Figure 4.4 and Table 4.4). However, contrary to cysteines involved in disulfide bridges, no complete oxidation of the cysteines involved in binding the heme as observed.

In our previously reported chemical selection strategy using CNBr, proteins were only cleaved C-terminal of methionine. Since methionine only accounts for 2.59% of the amino acids in *Shewanella oneidensis* MR-1 (test species used), the peptides generated after CNBr cleavage are relatively large. In order for peptides to be detected on a MALDI-TOF/TOF analyzer they need to be in the 1-5.5 kDa mass range. Our data showed that most of the succesfully *de novo* interpreted MS/MS spectra had a parent ion below 3.5 kDa. Using CNBr as cleavage reagent, 51% of the C-terminal peptides fall in the range of efficient detection and only 1 out of 3 proteins generates a C-terminal peptide in the *de novo* sequencing mass range. When KI is added to the CNBr cleavage reaction mixture, cleavage C-terminal of methionine and tryptophan is observed. Tryptophan accounts for 1.25% of the amino acids in *S. oneidensis*

MR-1. By cleaving C-terminal of tryptophan and methionine, 58% of the C-terminal peptides are detectable on MALDI-TOF/TOF MS and 42% of the proteins have a C-terminal peptide in the *de novo* sequencing mass range. As a proof-of-concept to demonstrate the improved proteome coverage, the chemical selection protocol was applied to the CNBr and KI cleaved proteins avidin and β-lactoglobulin. After partial ring opening the C-terminal peptide was clearly identified as a singlet amongst the internal peptide doublets, resp. 2131.14 and 2112.86. MS/MS analysis of the C-terminal peptide of avidin allowed to generate a C-terminal sequence tag.

4.3.6 Conclusions

Here we report on the refinement of a methodology to obtain chemical protein cleavage simultaneously after Met and Trp residues, combined with disulfide bond oxidation, in a single reaction. The need for reduction and alkylation of disulfide bridges requires laborious protocols that are using chemical cleavage methods in recent publications [35, 42, 57]. Given the interest that currently exists in increasing the sequence coverage during the study of proteins, and more specifically membrane proteins that are often difficult to study using enzymatic digests [58–60], the chemical cleavage at two residues could be of great interest. The potential proteome coverage of the previously described C-terminal sequencing techniques is also widened [44].

4.3.7 Acknowledgement

This work was supported by research grant G.0644.07N from FWO-Vlaanderen to B.D. and B.S. B.S. is indebted to a Postdoctoral fellow grant of the Fund for Scientific Research-Flanders (F.W.O.-Vlaanderen, Belgium). K.S. and P.M. are funded by a Ph.D. grant of the Institute for the promotion of Innovation through Science and Technology in Flanders (I.W.T.-Vlaanderen).

References

[1] Huang, S. and Huang, J. (1994) Cleavage of both tryptophanyl and methionyl peptide-bonds in proteins. *Journal of protein chemistry*, **13**, 450–451.

[2] Witkop, B. (1961) Non-enzymatic methods for the preferential and selective cleavage and modification of proteins. *Advances in protein chemistry*, **16**, 221–321.

[3] Spande, T. and Witkop, B. (1967) Determination of the tryptophan content of proteins with N-bromosuccinimide. *Methods in enzymology*, **11**, 498–506.

[4] Spande, T., Witkop, B., Degani, Y., and Patchornik, A. (1970) Selective cleavage and modification of peptides and proteins. *Advances in protein chemistry*, **24**, 97–260.

[5] Cohen, L. (1968) Group-specific reagents in protein chemistry. *Annual review of biochemistry*, **37**, 695–726.

[6] Patchornik, A., Lawson, W. B., Gross, E., and Witkop, B. (1960) The use of N-bromosuccinimide and N-bromoacetamide for the selective cleavage of C-tryptophyl peptide bonds in model peptides and glucagon. *Journal of the American chemical society*, **82**, 5923–5927.

[7] Schmir, G. L., Cohen, L. A., and Witkop, B. (1959) The oxidative cleavage of tyrosyl-peptide bonds. Cleavage of dipeptides and some properties of the resulting spirodienone-lactones. *Journal of the American chemical society*, **81**, 2228–2233.

[8] Shaltiel, S. and Patchornik, A. (1963) Cleavage of histidyl peptide bonds by N-bromosuccinimide. *Journal of the American chemical society*, **85**, 2799–2806.

[9] Omenn, G. S., Fontana, A., and Anfinsen, C. B. (1970) Modification of the single tryptophan residue of staphylococcal nuclease by a new mild oxidizing agent. *Journal of biological chemistry*, **245**, 1895–1902.

[10] Shechter, Y., Patchornik, A., and Burstein, Y. (1976) Selective chemical cleavage of tryptophanyl peptide bonds by oxidative chlorination with N-chlorosuccinimide. *Biochemistry*, **15**, 5071–5075.

[11] Funatsu, M., Green, N., and Witkop, B. (1964) Differential oxidation of protein-bound tryptophan and tyrosine by N-bromosuccinimide in urea solution. *Journal of the American chemical society*, **86**, 1846–1848.

[12] Fontana, A., Savige, W., Zambonin, M., and Birr, C. (1980) *Methods in peptide and protein sequence analysis*. Elsevier/North-Holland Biomedical Press, Amsterdam.

[13] Fontana, A., Dalzoppo, D., Grandi, C., and Zambonin, M. (1981) Chemical cleavage of tryptophanyl and tyrosyl peptide bonds via oxidative halogenation mediated by o-iodosobenzoic acid. *Biochemistry*, **20**, 6997–7004.

[14] Burstein, Y. and Patchornik, A. (1972) Selective chemical cleavage of tryptophanyl peptide bonds in peptides and proteins. *Biochemistry*, **11**, 4641–4650.

[15] Burstein, Y., Wilchek, M., and A, P. (1967) A selective chemical cleavage of tryptophyl peptide bonds. *Journal of the American chemical society*, **5**, P65.

[16] Junek, H., Kirk, K. L., and Cohen, L. A. (1969) Oxidative cleavage of tyrosyl-peptide bonds during iodination. *Biochemistry*, **8**, 1844–1848.

[17] Kingsbury, W. D. and Johnson, C. R. (1969) Oxidation of sulphides to sulphoxides with 1-chlorobenzotriazole. *Journal of the chemical society D: chemical communications*, pp. 365–365.

[18] Mahoney, W. C. and Hermodson, M. A. (1979) High-yield cleavage of tryptophanyl peptide bonds by o-iodosobenzoic acid. *Biochemistry*, **18**, 3810–3814.

[19] Alexander, N. M. (1974) Oxidative cleavage of tryptophanyl peptide bonds during chemical-and peroxidase-catalyzed iodinations. *Journal of biological chemistry*, **249**, 1946–1952.

[20] Lischwe, M. and Sung, M. (1977) Use of N-chlorosuccinimide/urea for the selective cleavage of tryptophanyl peptide bonds in proteins. Cytochrome c. *Journal of biological chemistry*, **252**, 4976–4980.

[21] Ozols, J. and Gerard, C. (1977) Cleavage of tryptophanyl peptide bonds in cytochrome b5 by cyanogen bromide. *Journal of biological chemistry*, **252**, 5986–5989.

[22] Caprioli, R. and CF, B. (1984) *GC-MN News*, **12**, 152.

[23] Vestling, M. M., Kelly, M. A., Fenselau, C., and Costello, C. E. (1994) Optimization by mass spectrometry of a tryptophan-specific protein cleavage reaction. *Rapid communications in mass spectrometry*, **8**, 786–790.

[24] Smith, L. M. and Kelleher, N. L. (2013) Proteoform: a single term describing protein complexity. *Nature methods*, **10**, 186–187, Top-Down Proteomics Consortium.

[25] Tran, J. C., et al. (2011) Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*, **480**, 254–258.

[26] Steen, H. and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nature reviews molecular cell biology*, **5**, 699–711.

[27] Paizs, B. and Suhai, S. (2005) Fragmentation pathways of protonated peptides. *Mass spectrometry reviews*, **24**, 508–548.

[28] Zhang, Y., Dewald, H. D., and Chen, H. (2011) Online mass spectrometric analysis of proteins/peptides following electrolytic cleavage of disulfide bonds. *Journal of proteome research*, **10**, 1293–1304.

[29] Basile, F. and Hauser, N. (2011) Rapid online non-enzymatic protein digestion combining microwave heating acid hydrolysis and electrochemical oxidation. *Analytical chemistry*, **83**, 359–367.

[30] Swatkoski, S., Gutierrez, P., Wynne, C., Petrov, A., Dinman, J. D., Edwards, N., and Fenselau, C. (2008) Evaluation of microwave-accelerated residue-specific acid cleavage for proteomic applications. *The journal of proteome research*, **7**, 579–586.

[31] Hua, L., Low, T. Y., and Sze, S. K. (2006) Microwave-assisted specific chemical digestion for rapid protein identification. *Proteomics*, **6**, 586–591.

[32] Gross, E. and Witkop, B. (1962) Non-enzymatic cleavage of peptide bonds: the methionine residues in bovine pancreatic ribonuclease. *Journal of biological chemistry*, **237**, 1856–1860.

[33] Ambler, R. (1965) Behaviour of peptides formed by cyanogen bromide cleavage of proteins. *Biochemical journal*, **96**, P32.

[34] Cook, A. D., Gray, R., Ramshaw, J., Mackay, I. R., and Rowley, M. J. (2004) Antibodies against the CB10 fragment of type II collagen in rheumatoid arthritis. *Arthritis research & therapy*, **6**, R477–83.

[35] Kuhn, K., Thompson, A., Prinz, T., Müller, J., Baumann, C., Schmidt, G., Neumann, T., and Hamon, C. (2003) Isolation of N-terminal protein sequence tags from cyanogen bromide cleaved proteins as a novel approach to investigate hydrophobic proteins. *Journal of proteome research*, **2**, 598–609.

[36] Katsumi, A., Tuley, E. A., Bodó, I., and Sadler, J. E. (2000) Localization of disulfide bonds in the cystine knot domain of human von Willebrand factor. *Journal of biological chemistry*, **275**, 25585–25594.

[37] Lisenbee, C. S., Dong, M., and Miller, L. J. (2005) Paired cysteine mutagenesis to establish the pattern of disulfide bonds in the functional intact secretin receptor. *Journal of biological chemistry*, **280**, 12330–12338.

[38] Goodlett, D. R., Armstrong, F. B., Creech, R. J., and van Breemen, R. B. (1990) Formylated peptides from cyanogen bromide digests identified by fast atom bombardment mass spectrometry. *Analytical biochemistry*, **186**, 116–120.

[39] Morrison, J., Fidge, N., and Grego, B. (1990) Studies on the formation, separation, and characterization of cyanogen bromide fragments of human AI apolipoprotein. *Analytical biochemistry*, **186**, 145–152.

[40] Kaiser, R. and Metzka, L. (1999) Enhancement of cyanogen bromide cleavage yields for methionyl-serine and methionyl-threonine peptide bonds. *Analytical biochemistry*, **266**, 1–8.

[41] Wang, Z., Hilder, T. L., van der Drift, K., Sloan, J., and Wee, K. (2013) Structural characterization of recombinant α-1-antitrypsin expressed in a human cell line. *Analytical biochemistry*, **437**, 20 – 28.

[42] Samyn, B., Sergeant, K., Castanheira, P., Faro, C., and Van Beeumen, J. (2005) A new method for C-terminal sequence analysis in the proteomic era. *Nature methods*, **2**, 193–200.

[43] Moerman, P., Sergeant, K., Debyser, G., Devreese, B., and Samyn, B. (2010) A new chemical approach to differentiate carboxy terminal peptide fragments in cyanogen bromide digests of proteins. *Journal of proteomics*, **73**, 1454–1460.

[44] Moerman, P., Sergeant, K., Debyser, G., Timperman, I., Devreese, B., and Samyn, B. (2014) Automation of C-terminal sequence analysis of 2D-PAGE separated proteins. *EuPA open proteomics*, **3**, 250–261.

[45] Savige, W. E. and Fontana, A. (1977) Modification of tryptophan to oxindolylalanine by dimethyl sulfoxide-hydrochloric acid. *Methods in enzymology*, **47**, 442.

[46] Taylor, S. W., Fahy, E., Murray, J., Capaldi, R. A., and Ghosh, S. S. (2003) Oxidative post-translational modification of tryptophan residues in cardiac mitochondrial proteins. *Journal of biological chemistry*, **278**, 19587–19590.

[47] Permentier, H. P. and Bruins, A. P. (2004) Electrochemical oxidation and cleavage of proteins with on-line mass spectrometric detection: development of an instrumental alternative to enzymatic protein digestion. *Journal of the american society for mass spectrometry*, **15**, 1707–1716.

[48] Rosa, J. C., Greene, L. J., De Oliveira, P. S. L., Garratt, R., Beltramini, L., Resing, K., and Roque-Barreira, M.-C. (1999) KM+, a mannose-binding lectin from *Artocarpus integrifolia*: Amino acid sequence, predicted tertiary structure, carbohydrate recognition, and analysis of the β-prism fold. *Protein science*, **8**, 13–24.

[49] Smith, B. and Walker, J. (2002) *The protein protocols handbook*, vol. 2. Humana Press Inc.: Totowa.

[50] Lang, R. (1925) Über neue jodometrische Methoden, die auf der Bildung und Messung von Jodcyanid beruhen. IV. *Zeitschrift für anorganische und allgemeine Chemie*, **144**, 75–84.

[51] Ravindranath, K. and Patel, C. (1968) Volumetric determination of tetraalkyl thiuram disulphides. *Fresenius' Zeitschrift für Analytische Chemie*, **238**, 276–278.

[52] Huang, H. V., Bond, M. W., Hunkapiller, M. W., and Hood, L. E. (1983) Cleavage at tryptophanyl residues with dimethyl sulfoxide-hydrochloric acid and cyanogen bromide. *Methods in enzymology*, **91**, 318–324.

[53] Shinohara, K. (1932) Oxidation of cystine by iodine in aqueous medium. *Journal of biological chemistry*, **96**, 285–297.

[54] Crimmins, D. L., McCourt, D. W., Thoma, R. S., Scott, M. G., Macke, K., and Schwartz, B. D. (1990) In situ chemical cleavage of proteins immobilized to glass-fiber and polyvinylidenedifluoride membranes: Cleavage at tryptophan residues with 2-(2 -nitrophenylsulfenyl)-3-methyl-3 -bromoindolenine to obtain internal amino acid sequence. *Analytical biochemistry*, **187**, 27–38.

[55] Patrie, K. M., Botelho, M. J., Franklin, K., and Chiu, I.-M. (1999) Site-directed mutagenesis and molecular modeling identify a crucial amino acid in specifying the heparin affinity of FGF-1. *Biochemistry*, **38**, 9264–9272.

[56] Lederer, F. and Tarin, J. (1971) Chemical modification of the thioether bridges in cytochrome c. *European journal of biochemistry*, **20**, 482–487.

[57] Prinz, T., Müller, J., Kuhn, K., Schäfer, J., Thompson, A., Schwarz, J., and Hamon, C. (2004) Characterization of low abundant membrane proteins using the protein sequence tag technology. *Journal of proteome research*, **3**, 1073–1081.

[58] van Montfort, B. A., Doeven, M. K., Canas, B., Veenhoff, L. M., Poolman, B., and Robillard, G. T. (2002) Combined in-gel tryptic digestion and CNBr cleavage for the generation of peptide maps of an integral membrane protein with MALDI-TOF mass spectrometry. *Biochimica et biophysica acta - Bioenergetics*, **1555**, 111–115.

[59] Laugesen, S. and Roepstorff, P. (2003) Combination of two matrices results in improved performance of MALDI MS for peptide mass mapping and protein analysis. *Journal of the american society for mass spectrometry*, **14**, 992–1002.

[60] Han, S.-Y. and Kim, Y.-A. (2004) Recent development of peptide coupling reagents in organic synthesis. *Tetrahedron*, **60**, 2447–2467.

Chapter 5

C-terminal selection by piperazine labelling

5.1 Introduction

The methodologies described in the previous chapters are only applicable to proteins separated by gel electrophoresis. In the time frame of our project, 2D-PAGE separation in proteomics was increasingly overtaken by LC-MS based methods. Although 2D-PAGE still has a place in proteomics, the high reproducibility, resolution and ability to analyze hydrophobic proteins and proteins with high and low pI values, make shotgun LC-MS a more powerful platform [1–3]. In addition, the low ionization and fragmentation efficiency of large CNBr-derived C-terminal peptides (lacking a terminal Lys or Arg) by MALDI-TOF/TOF MS triggered us to develop a completely new strategy to select for these peptides in a form that can be consequently analyzed on any (LC-) MS platform using all available fragmentation techniques with high sensitivity. The most cited methods for LC-MS based C-terminal peptide identification and characterization are COFRADIC [4] and C-TAILS [5]. However, both methods did not truly make it as standard methods, due to the complexity and workload of the former, and the in-house synthesized coupling polymer of the latter. Therefore, we aimed to develop a more simple method based on modification and enrichment of C-terminal peptides.

Unlike internal tryptic peptides, C-terminal peptides rarely end on a basic residue. Since the ionization efficiency in MALDI is dependent on the hydrophobicity and gas-phase basicity of the analyte, C-terminal peptides often remain undetected in a classical MALDI-TOF peptide mass fingerprint. Similarly, maximal ESI signal responses in the positive ionization mode are observed at solvent pH values three or more units lower than isoelectric point of the analyte [6]. Several modifications have been presented to improve the hydrophobicity, pI and gas-phase basicity of C-terminal peptides and peptides containing a negatively charged post-translational modification, such as phosphorylation [7–9]. Yang *et al.* used 1-(2-pyrimidyl) piperazine to

modify all carboxyl groups using a carbodiimide mediated coupling at nearly 100% yield, increasing the ionization efficiency for phosphopeptides 50-100 times in MALDI MS. The same method allowed to determine phosphorylation sites using ESI-ETD-MS [10, 11] (Figure 5.1). In an additional study 4 piperazines, i.e. (1-(2-pyridyl)piperazine, 1-(2-pyrimidyl)piperazine, 1-(4-pyridyl)piperazine and 1-(1-methyl-4-piperidinyl)piperazine), were compared to evaluate their coupling efficiency and improvement of hydrophobicity and gas-phase basicity of the modified peptides. (1-(2-pyridyl)piperazine and 1-(2-pyrimidyl)piperazine were found to be preferable to increase the peptide signals on MALDI-TOF MS [12] (Figure 5.2).

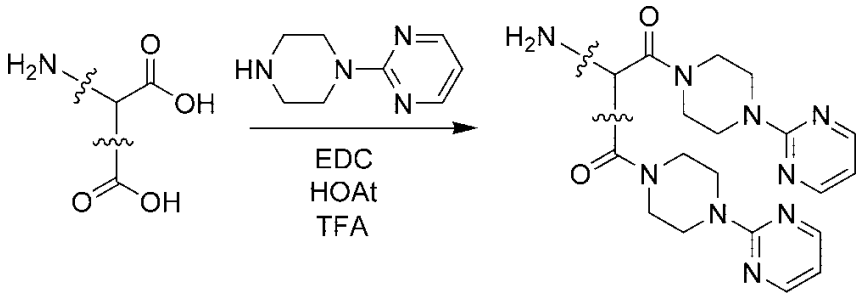


Figure 5.1: Synthesis of 1-(2-pyrimidyl)piperazine derivatized proteins. EDC = 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide hydrochloride; HOAt = 1-hydroxy-7-azabenzotriazole; TFA = trifluoroacetic acid [10].

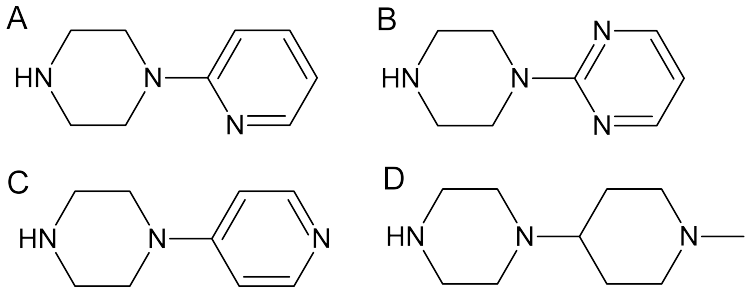


Figure 5.2: Different piperazines used; 1-(2-pyridyl)piperazine (Panel A), 1-(2-pyrimidyl)piperazine (Panel B), 1-(4-pyridyl)piperazine (Panel C) and 1-(1-methyl-4-piperidinyl)piperazine (Panel D).

Here we have adapted this derivatization method towards a new technique to probe protein C-termini that has potential to be applied both for in-gel and in-solution separated proteins. We used the carbodiimide coupling reaction to modify all side chain carboxyl groups as well as the C-terminus (Figure 5.1). After this modification step, the proteins are digested using trypsin. The newly formed carboxyl groups are then captured onto a COOH-coupling matrix using the same carbodiimide coupling chemistry as used in the first step, leaving the C-terminal peptide in solution ready to be analyzed and sequenced using any type of mass spectrometric setup (Figure

5.3). The piperazine derivatization does not only serve to improve the hydrophobicity and gas-phase basicity of the C-terminal peptide, but the specific mass tag (+146.19 Da) can be used as a positive control to distinguish it from internal peptides. The analysis on an LC-MS system allows the use of other fragmentation techniques than MALDI (CID) MS. As demonstrated by ETD on phosphorylated peptides by Zhang, modified C-terminal peptides should generate a sequence-independent fragmentation behavior over the entire peptide backbone, allowing their identification by tandem mass spectrometry [11]. We selected coupling chemistry that is well known in peptide synthesis and bioconjugate production to prepare the piperazine derivatized peptides. 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide hydrochloride (EDC) was used as coupling reagent in combination with 1-hydroxy-7-azabenzotriazole (HOAt) to avoid the O-acyl urea rearrangement [13] and to activate the intermediate for amide formation (Figures 5.4, 5.5, 5.6). Here we present the initial results of the development and optimization of this method, and discuss strategies for further optimization.

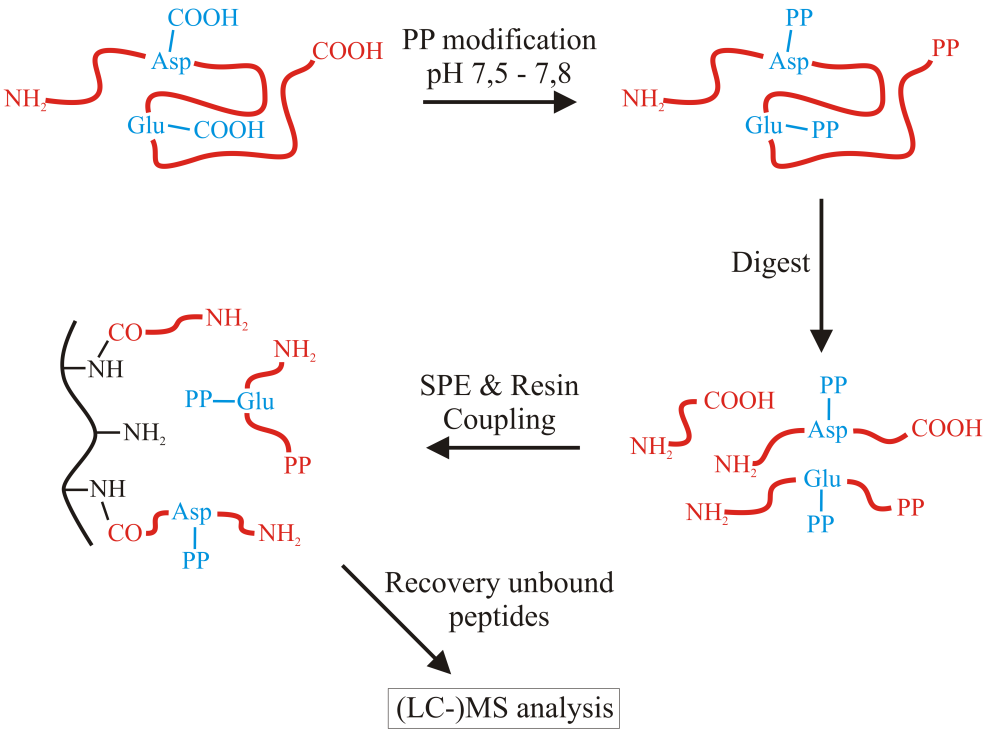


Figure 5.3: Schematic representation of the different steps in the C-terminal sequencing method using piperazine modification. Step 1: All carboxyl functions in the protein are modified. Step 2: The modified proteins are cleaved. Step 3: The new generated carboxyl functions are bound to a COOH binding matrix using the same coupling chemistry as in step 1. Step 4: Modified C-terminal peptides with enhanced ionization characteristics can be analyzed using any kind of (LC-)MS setup.

5.1.1 EDC coupling

Carbodiimides are dehydration agents often used to activate carboxylic acids towards amide or ester formation. Sheehan reported the first formation of peptide bonds using carbodiimides [14]. Several carbodiimides have been used in solid-phase peptide synthesis; N,N'-Dicyclohexylcarbodiimide (DCC), N,N'-Diisopropylcarbodiimide (DIC), 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide hydrochloride (EDC) [15–17]. Since DCC is insoluble in water and DIC is very toxic, EDC is currently the prefered coupling reagent in amide formation.

The first step of the amide formation reaction mechanism involves the intermediate formation of the activated O-acylisourea derivative of the carbodiimide. A subsequent nucleophilic attack by the primary nitrogen of the amino compound brings about the formation of the wanted amide linkage, with the release of the soluble substituted urea. Alternatively the O-acylisourea can also be attacked by a second carboxylate to generate an anhydride, which can then be attacked by the amine, producing the amide and regenerating one of the carboxylates [18–21]. The formation of O-acylurea occurs optimally at pH 4-5; the primary amino group of the nucleophile is predominantly protonated at this low pH and therefore rather unreactive [22]. The intermediate has an extremely short half-life and rapidly undergoes hydrolysis or rearranges to an N-acylurea adduct [23]. Adding 1-hydroxy-7-azabenzotriazole (HOAt) to the reaction mixture minimizes the side reaction. It reacts with the instable O-acylisourea intermediate to form a more stable insoluble activated ester that reacts with amines at ambient temperatures to generate amides [13]. In an additional unwanted head-to-tail side reaction the peptide N-terminus competes with the piperazine amine to form cross-linked or circular peptides.

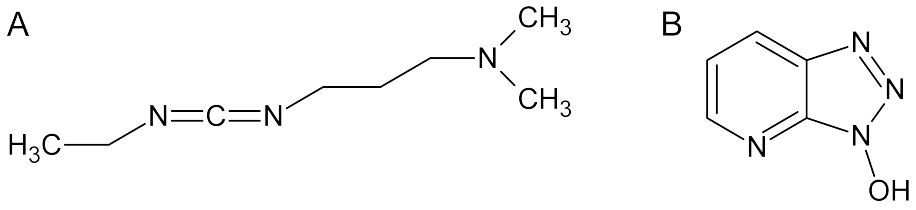


Figure 5.4: Structure of 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide hydrochloride (ECD) (Panel A) and 1-hydroxy-7-azabenzotriazole (HOAt) (Panel B).

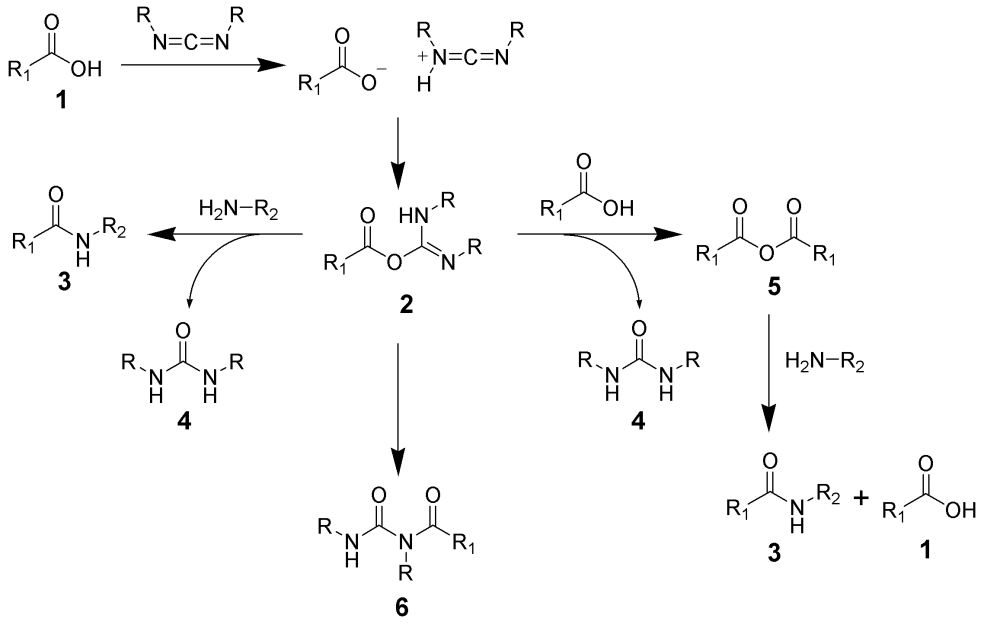


Figure 5.5: Reaction mechanism of amide bond formation using carbodiimide coupling. The first reaction steps an O-acylisourea-derivative is formed. This intermediate can form the amide bond by reacting with the amino group directly (left pathway) or by first reacting with a second carboxylate group forming an anhydride intermediate. The O-acylisourea intermediate can also rearrange and form an unwanted, stable N-acylurea adduct. 1 = Carboxyl group, 2 = O-acylisourea, 3 = amide, 4= urea, 5= acid anhydride, 6 = N-acylurea [14, 24].

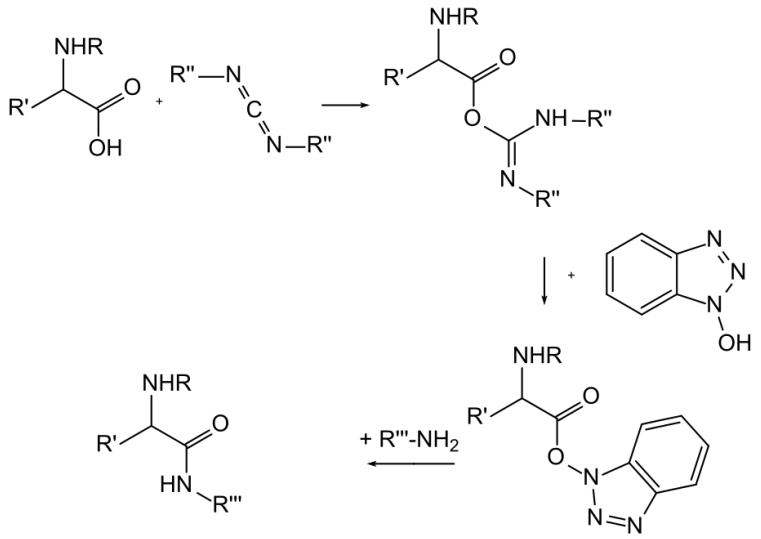


Figure 5.6: Reaction mechanism of active-ester formation using 1-hydroxy-7-azabenzotriazole (HOAt). HOAt can be added to the carbodiimide coupling reaction to limit the formation of the stable N-acylurea adduct. HOAt rapidly reacts with the O-acylisourea to form a more stable activated ester, that in turn reacts with an amine to generate the wanted amide bond [13, 25].

5.2 Materials and Methods

5.2.1 Materials

Angiotensin, bradykinin, yeast alcohol dehydrogenase (gel filtration purity) and horse heart cytochrome c (purity 99%) were purchased from Sigma (Bornem, Belgium). A part (N35 to Arg277) of the XcpQ subunit of the Type II secretion system protein D of *Pseudomonas aeruginosa* (Uniprot: P35818) was expressed and purified in house [26]. Stock solutions (0,5 nmol/μl) were prepared for all proteins and further diluted prior to use. HPLC grade acetonitrile (ACN) was obtained from BioSolve (Valkenswaard, The Netherlands). Trifluoroacetic acid (TFA) (purity >99.9%) was from Applied Biosystems (Foster city, CA, U.S.A.). Ammonium hydrogen carbonate (NH_4HCO_3) (purity >99.5%), dimethylformamide (DMF) (purity >99,8%), 1-(2-pyridyl)piperazine (purity >99%), 1-(2-pyrimidyl)piperazine (purity >98%), 1-(3-dimethylaminopropyl)-3-ethylcarbodiimide HCl (purity >99%), 0,6 M 1-hydroxy-7-azabenzotriazole (HOAt) in DMF (purity >98%) and α-cyano-4-hydroxycinnamic acid were purchased from Sigma. Immobilized diaminodipropylamine CarboxyLink coupling gel (Figure 5.7) was purchased from Thermo scientific (San Jose, CA, USA). Water was purified using a Millipore MilliQ water filtration system (Billerica, MA, USA).

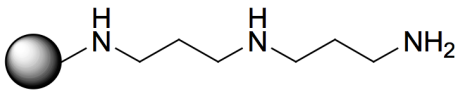


Figure 5.7: Structure of the immobilized diaminopropylamine coupling gel.

5.2.2 Production of piperazine derivatized peptides or proteins

A modification reagent stock solution was prepared containing 25 ml MQ; 1.5ml (2mg/ml) HOAt in DMF; 2ml (2mg/ml) EDC in DMF; 3ml 0.5% 1-(2-pyrimidyl)piperazine (PP) in DMF. TFA was added to the mixture until a pH between 7.5 and 7.8 was reached. 20 μl of the PP reaction mixture was added to 5 μl peptide/protein (10 pmol/μl), vortexed and incubated at room temperature for 1 hour. After incubation the samples are dried in the SpeedVac (Thermo Savant) to stop the reaction. The dried modified peptides were redissolved in 50% ACN/0.1% TFA and spotted on a MALDI plate. The modified proteins were then digested using trypsin. Trypsin is active between pH 7 and 9, which allows to perform the cleavage outside the EDC activity range and thus prevents immediate modification of the newly formed carboxyl functions. The digested protein samples were cleaned up using ZipTip solid phase extraction prior to analysis. MALDI-TOF MS sample preparation analysis was performed as previously described [27].

5.2.3 Coupling to CarboxyLink gel

The coupling of the peptides to the immobilized diaminodipropylamine gel was performed as described in the instructions. A disposable polypropylene column was filled with 4 ml of the CarboxyLink gel slurry and equilibrated with 5 column volumes of coupling solution. The sample (2 ml in coupling solution) was added to the gel slurry and gently mixed end-over-end for several minutes. 60 mg of EDC dissolved in 0.5 ml coupling solution was added to the gel and gently mixed for 3 hours at room temperature. After the gel had settled, the liquid phase and one column volume of washing solution (50% ACN/0.1% TFA) were collected containing the non-bound peptide fraction. After further washing, the column was loaded with PBS buffer containing 0.05% of sodium azide and stored until further use at 4 °C. We used 50% DMF in water as coupling solution and 50% ACN/0.1% TFA as wash solution. According to the supplier, anything but phosphate, acetate, tris and glycine or thiol containing buffers can be used since they are known to react with the O-acylintermediate and/or inactivate EDC.

Remarks:

- EDC is moisture sensitive and hydrolyzes quickly when dissolved in aqueous buffers. The stock solution needs to be made fresh every time and the EDC reagent needs to be stored in a tightly sealed vial with dessicant.
- Although the active pH range of EDC is often reported to be between pH 4 and 6, the best coupling results were obtained in the narrow pH range 7.5-7.8.
- All DMF needs to be evaporated before samples are redissolved and mixed with α-cyano-4-hydroxycinnamic acid matrix mixture. Otherwise matrix peaks will appear in the higher mass range disturbing MS interpretation.

5.3 Results and Discussion

5.3.1 Piperazine modification of test peptides

Carbodiimide mediated coupling of piperazines to carboxyl groups is not specific, all side chain carboxyl groups and the C-terminus of the protein are thus expected to be modified. The reaction was optimized on two test peptides, bradykinin and angiotensin, the latter containing an N-terminal Asp side chain carboxyl group (Figures 5.8, 5.9). Analysis of samples taken 10 minutes after initiation of the reaction indicate that modification of both sites occurs simultaneously. Indeed, MS/MS analysis of the singly modified angiotensin (1442.88 Da) indicates that a mix of the C-terminal and side chain modified peptides is obtained (Figure 5.10 and Table 5.1). In a minor, unwanted head-to-tail side reaction the amino terminus of the peptides competes with the PP reagents to couple to the activated-ester intermediate,

forming cross-links between peptides (Figure 5.8). The cross-linking of both peptides was mainly observed at higher concentrations of peptide, indicating that the phenomenon can be avoided by reducing the peptide to PP reagent ratio. During these initial experiments, we compared the mass spectrometric response between peptides modified with 1-(2-pyridyl)piperazine and 1-(2-pyrimidyl)piperazine. The results were comparable, but 1-(2-pyrimidyl)piperazine was chosen as reagent for further experiments.

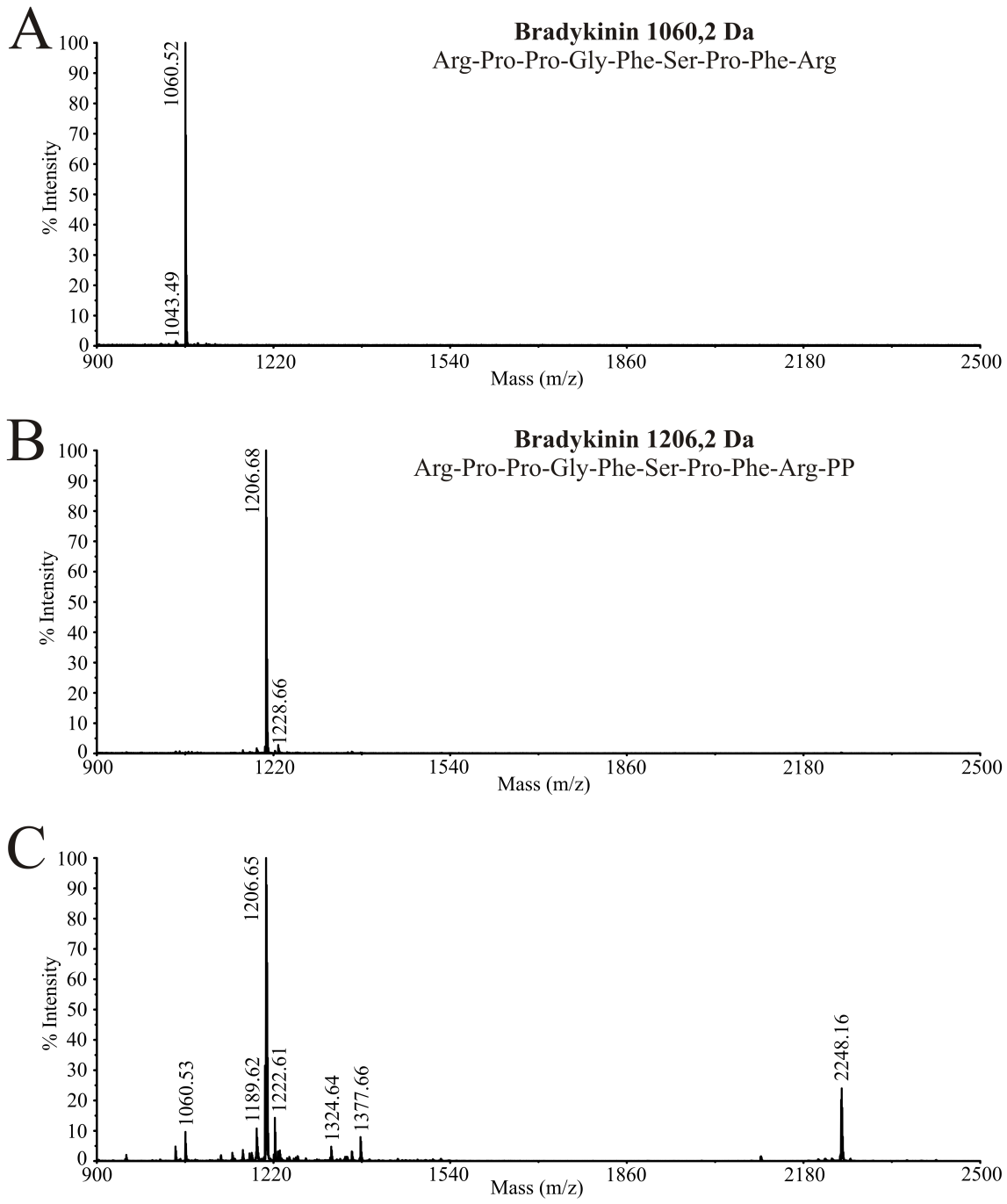


Figure 5.8: Piperazine derivatization of bradykinin. Panel A: MALDI MS spectrum of bradykinin (theoretical mass : 1060.57 Da) before incubation with PP reaction mixture. Panel B: MALDI MS spectrum of bradykinin containing a C-terminal PP modification (theoretical mass : 1206.76 Da). Panel C: At higher peptide/PP reagent ratio (4 times), the peptide with mass 2248.31 is formed, due to a head-to-tail side reaction.

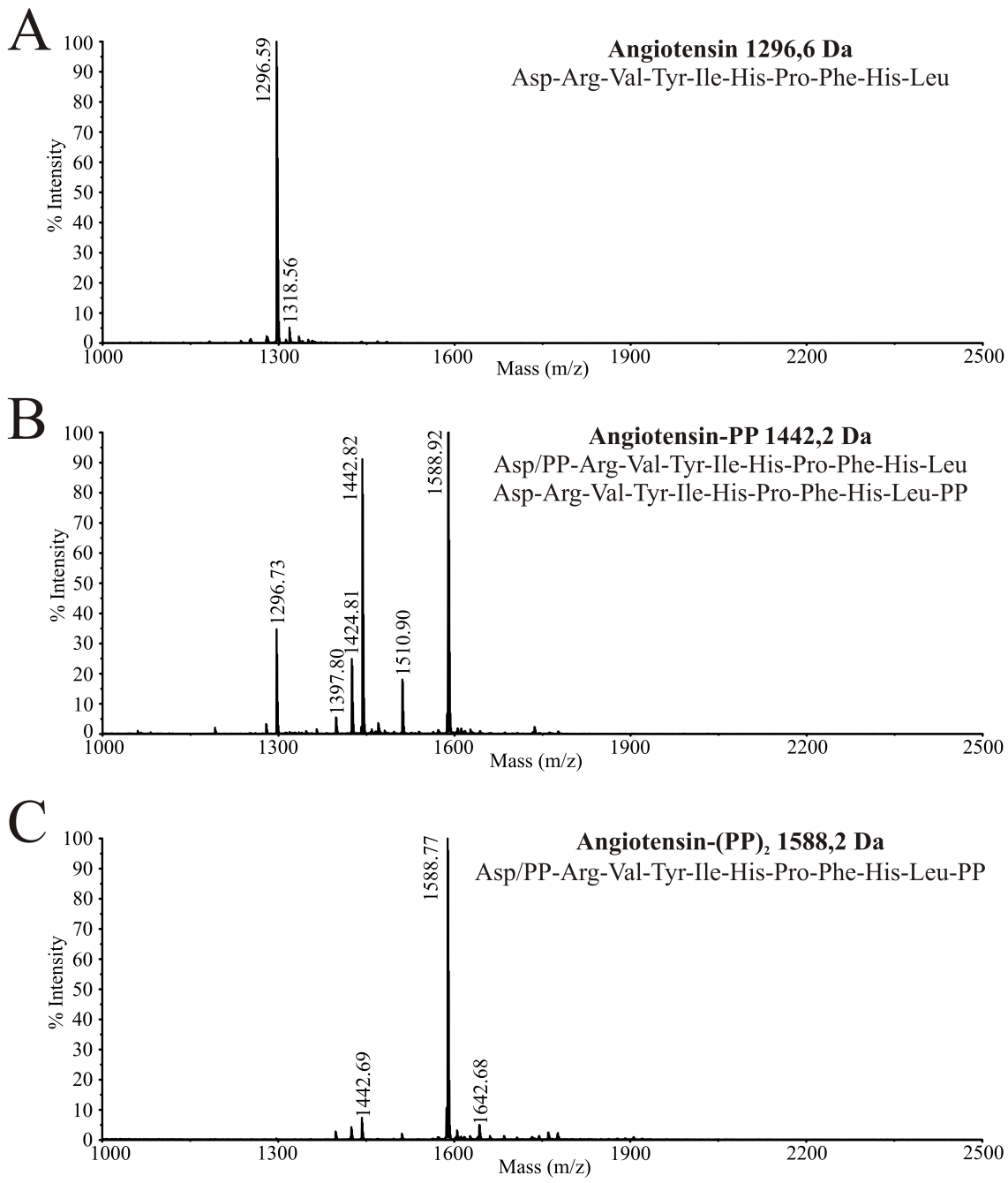


Figure 5.9: Piperazine derivatization of angiotensin. Panel A: MALDI MS spectrum of angiotensin (1296.69 Da) prior to incubation with PP reaction mixture. Panel B: MALDI MS spectrum of angiotensin 10 minutes after initiation of the reaction. The peptide with mass 1442.88 Da contains one PP modification, either at the C-terminus or on the Asp side chain. Panel C: MALDI MS spectrum of angiotensin containing two PP modifications (1588.07 Da).

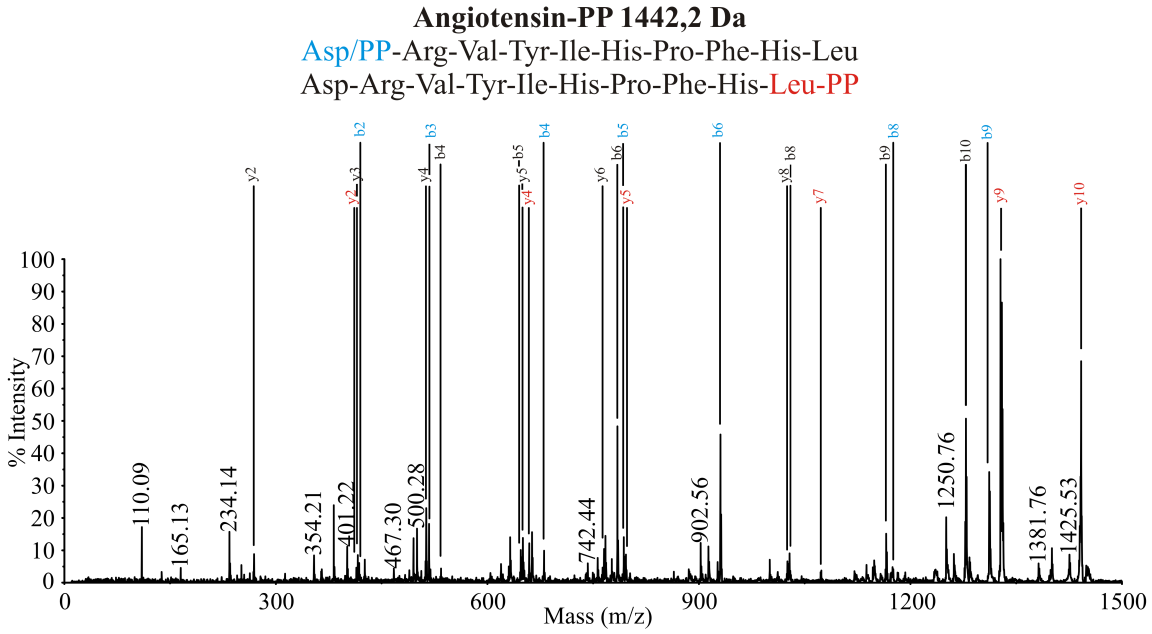


Figure 5.10: MS/MS spectrum of singly modified angiotensin, mass 1442.88 Da. The fragmentat ions corresponding to the unmodified peptide are marked in black. The C-terminal modified peptide y-ions are 146.19 Da heavier and are marked in red, the Asp side chain (N-terminal amino acid) modified peptides are marked in blue. Both species are observed simultaneously, showing that the modification does not discriminate the side chain and C-terminal carboxyl groups.

Table 5.1: Overview of fragment ions of singly modified angiotensin peptide.

b-ion	+146.19	262.13	**418.23**	**517.30**	**680.36**	**793.45**	**930.51**	1027.56	**1174.63**	**1311.69**	1424.77	
b-ion		116.03	272.14	371.20	**534.27**	**647.35**	**784.41**	881.46	**1028.53**	**1165.59**	**1278.67**	
	NH_2	**D**	**R**	**V**	**Y**	**I**	**H**	**P**	**F**	**H**	**L**	COOH
y-ion		1296.69	1181.66	**1025.56**	926.49	**763.43**	**650.35**	**513.29**	**416.23**	**269.16**	132.10	
y-ion	+146.19	**1442.78**	**1327.75**	1171.65	**1072.58**	909.52	**796.44**	**659.38**	562.33	**415.26**	278.20	

List of b- and y-ions of PP modified and unmodified angiotensin peptides. Observed ions in Figure 5.10 are marked in bold, C-terminal modified y-ions are shown in red, PP-modified b-ions are marked in blue.

5.3.2 Evaluation of the method on a test protein

We first evaluated the PP modification reaction by MALDI MS analysis of the modified and unmodified protein digest. The spectra show that the peptides are present in multiple forms in the reaction mixture and that the reaction was incomplete after 1 h of incubation (Figure 5.11). Although some peptides are expected to have three derivatized carboxyl groups, all of them have one or more free carboxyl groups left (Table 5.2). Given the large number of Asp and Glu present in the peptides and the observation how few modifications each peptide has undergone, the overall yield of the reaction must be around 10%. The C-terminal peptide (1512.84 Da) is observed

as unmodified, singly modified (1658.94 Da) and doubly modified species (1805.06 Da), in the latter case carrying a PP modification on the only internal carboxylgroup and on the C-terminus.

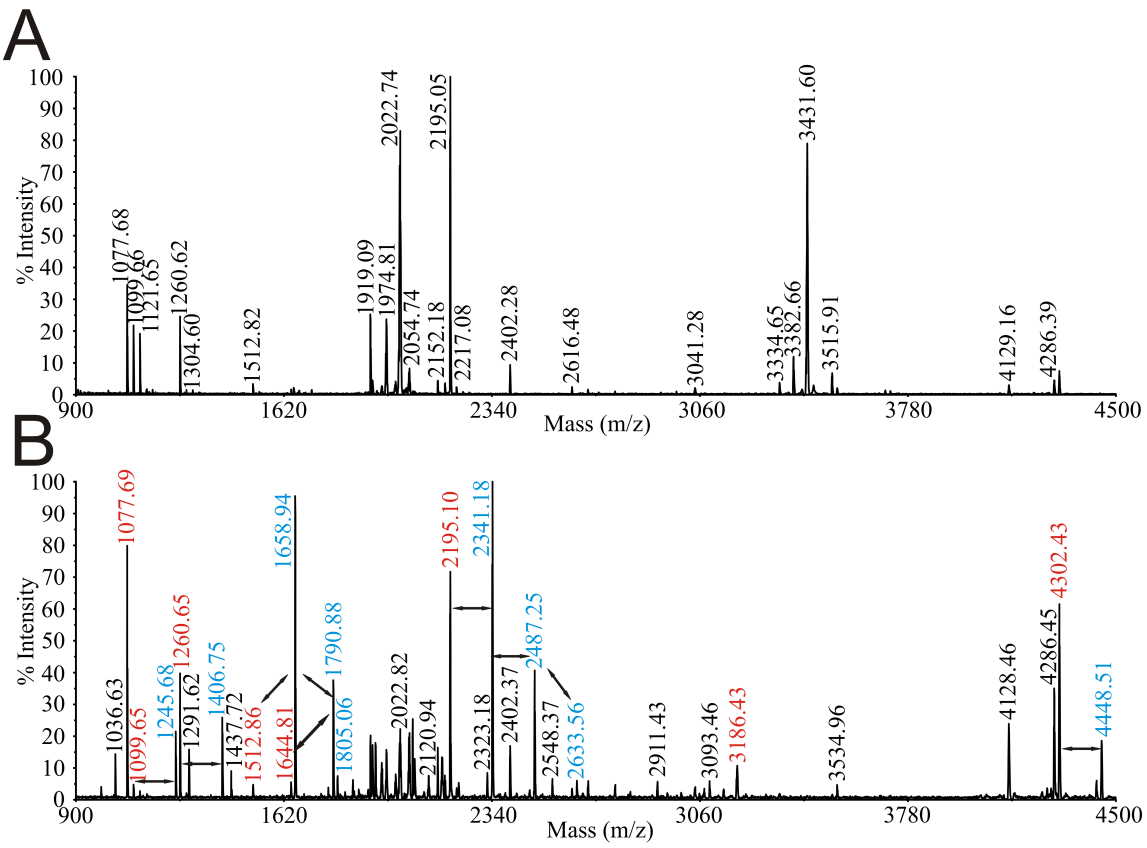


Figure 5.11: Evaluation of the piperazine derivatization of type II secretion system D (XcpQ, N35 to Arg277) of *Pseudomonas aeruginosa* by tryptic peptide mass fingerprinting. Panel A: MALDI MS spectrum of the tryptic digested protein without incubation with PP reaction mixture. Panel B: MALDI MS spectrum of the tryptic digested protein after incubation with PP reaction mixture. The arrows indicate the mass shifts of peptides as a result of the derivatization. For several peptides that have negatively charged side chains (e.g. 1260.65 and 2195.11), incomplete modification is shown. The C-terminal peptide (1512.84 Da) is observed as unmodified, singly modified (1658.94 Da) and doubly modified peptide (1805.06 Da).

The derivatization reaction at the protein level can be sterically hindered as a result of the tertiary and quaternary protein structure. Higher concentrations of DMF in the reaction mixture will in most cases destabilize the folded protein, allowing the reagent to access the internal carboxyl groups [28]. However, our results suggest that a stronger chaotrope or addition of detergent might be needed to (partially) unfold the protein and make all carboxylgroups accessible. Alternatively another well-known protein modification reaction can be used to permanently destabilize the protein by chemically altering some amino acids. O-methylisourea

modifies all lysines to homoarginine residues (Δ mass = 42 Da) [29, 30]. Homoarginine has a higher pK_a and increases the ionization efficiency of the peptide. The modification also increases the average mass of the tryptic peptides, as trypsin is unable to cleave C-terminal of homoarginine. Due to the reaction conditions, 65 °C at pH 10.6 in a 7 N NH_4OH solution, the samples will need to be desalted to lower the buffer pH prior to carbodiimide coupling. Additional sample cleanup can cause sample contamination or sample loss and should be limited. Alternatively, proteins can be reduced and alkylated using dithiothreitol and iodoacetamide in guanidium HCl/Tris buffer.

Table 5.2: List of trypic peptides of *Pseudomonas aeruginosa* type II secretion system (Xcp) observed

Peptide mass	**Amino acid nr.**	**Sequence**	**PP modified peptides**
4302.43	113-152	VIQVQQSPVSELIPLIRPLVPQYGH-LAAVPSANALIISDR	4448.51
3168.50	170-197	GSHDYSVINL**R**YGWVMDAAEVLN-NAMSR	
2195.09	29-47	EFIDQISEITGETFVVDPR	2341.18 - 2487.25 - 2633.56
1644.79	97-112	TEAGGGQSAPD**R**LETR	1790.88
1512.84	229-242	LVQLAQSLDTPTAR	1658.94 - 1805.06
1260.63	170-180	GSHDYSVINLR	1406.75
1099.59	202-213	GAAGAQVIADAR	1245.68
1077.68	217-226	LIILGPPQAR	

The observed masses of the unmodified peptides are in red, the masses of the PP modified peptides are in blue, missed cleavages are indicated in bold.

The most important carboxyl function to be labeled during the initial derivatization is the C-terminal one. This label serves as a mass tag for positive identification during MS/MS. Due to solvation kinetics and the importance of the terminal regions in protein binding in general and in post-translational modification, most termini are positioned at the solvated surface of the protein, well within reach of the reagents and should be modified [31–33].

5.3.3 Evaluation of capturing test peptides to the coupling matrix

A mixture of PP modified bradykinin and unmodified angiotensin was coupled to COOH-binding matrix, simulating a very simple digest mixture of a modified protein. After the reaction, the unbound fraction was collected, dried, redissolved in 50% ACN/0.1% TFA and analyzed on MALDI-TOF MS. The recovery was very low, although 4 nmol of each peptide was applied to the matrix. Besides modified bradykinin, also partially modified angiotensin was observed, showing that not all PP reagent had been removed during the sample cleanup of the bradykinin

mixture using ZipTips, and that coupling to the matrix was not 100% efficient. Since the gel has an activation level of 16-20 mol amine/mL of gel, the failure to bind all carboxylgroups was not due to overloading. In the current protocol, the composition and pH range of the coupling solution are similar to the one used to couple the PP reagent. Other coupling buffer compositions with different pH can be evaluated. The incomplete coupling of the internal peptides to the resin results in more complex MS spectra, but the C-terminal peptide should still be identified due to the specific PP mass shift in MS spectra.

5.4 Conclusions and future perspectives

We here presented a new strategy for C-terminal sequence analysis. The aim of this project was to develop a technique to positively select the C-terminal peptide while improving its ionization and fragmentation behavior. The protocol has to be applicable to all (LC-)MS platforms and cover a large area of the proteome by being compatible with different kinds, or combinations of proteases. Only the initial steps have been optimized, but the PP modification technique has the potential to fulfill these conditions.

So far only the coupling of the PP reagent to peptides is optimized, all other steps of the protocol still need further optimization. Several options are still available to counter some of the observed problems; the PP-modification of internal carboxylgroups in proteins can be improved by adding chaotropes to the reaction mixture, different buffers can be tested during resin coupling, miniaturization of the coupling to resin, different solid phase extraction resins and ultrafiltration tools can be compared during sample cleanup.

References

[1] Rogowska-Wrzesinska, A., Le Bihan, M.-C., Thaysen-Andersen, M., and Roepstorff, P. (2013) 2D gels still have a niche in proteomics. *Journal of proteomics*, **88**, 4–13.

[2] Braun, R. J., Kinkl, N., Beer, M., and Ueffing, M. (2007) Two-dimensional electrophoresis of membrane proteins. *Analytical and bioanalytical chemistry*, **389**, 1033–1045.

[3] Bunai, K. and Yamane, K. (2005) Effectiveness and limitation of two-dimensional gel electrophoresis in bacterial membrane protein proteomics and perspectives. *Journal of chromatography B*, **815**, 227–236.

[4] Van Damme, P., Staes, A., Bronsoms, S., Helsens, K., Colaert, N., Timmerman, E., Aviles, F. X., Vandekerckhove, J., and Gevaert, K. (2010) Complementary positional proteomics for screening substrates of endo- and exoproteases. *Nature methods*, **7**, 512–515.

[5] Schilling, O., Barre, O., Huesgen, P. F., and Overall, C. M. (2010) Proteome-wide analysis of protein carboxy termini: C terminomics. *Nature methods*, **7**, 508–U33.

[6] Pan, P., Gunawardena, H. P., Xia, Y., and McLuckey, S. A. (2004) Nanoelectrospray ionization of protein mixtures: solution pH and protein pI. *Analytical chemistry*, **76**, 1165–1174.

[7] Nakazawa, T., Yamaguchi, M., Nishida, K., Kuyama, H., Obama, T., Ando, E., Okamura, T., Ueyama, N., Tanaka, K., and Norioka, S. (2004) Enhanced responses in matrix-assisted laser desorption/ionization mass spectrometry of peptides derivatized with arginine via a C-terminal oxazolone. *Rapid communications in mass spectrometry*, **18**, 799–807.

[8] Yamaguchi, M., et al. (2006) Enhancement of MALDI-MS spectra of C-terminal peptides by the modification of proteins via an active ester generated *in situ* from an oxazolone. *Analytical chemistry*, **78**, 7861–7869.

[9] Nakajima, C., Kuyama, H., Nakazawa, T., and Nishimura, O. (2012) C-terminal sequencing of protein by MALDI mass spectrometry through the specific derivatization of the α-carboxyl group with 3-aminopropyltris-(2, 4, 6-trimethoxyphenyl) phosphonium bromide. *Analytical and bioanalytical chemistry*, **404**, 125–132.

[10] Xu, Y., Zhang, L., Lu, H., and Yang, P. (2008) Mass spectrometry analysis of phosphopeptides after peptide carboxy group derivatization. *Analytical chemistry*, **80**, 8324–8328.

[11] Zhang, L., Xu, Y., Lu, H., and Yang, P. (2009) Carboxy group derivatization for enhanced electron-transfer dissociation mass spectrometric analysis of phosphopeptides. *Proteomics*, **9**, 4093–4097.

[12] Qiao, X., Sun, L., Chen, L., Zhou, Y., Yang, K., Liang, Z., Zhang, L., and Zhang, Y. (2011) Piperazines for peptide carboxyl group derivatization: effect of derivatization reagents and properties of peptides on signal enhancement in matrix-assisted laser desorption/ionization mass spectrometry. *Rapid communications in mass spectrometry*, **25**, 639–646.

[13] Carpino, L. A. (1993) 1-hydroxy-7-azabenzotriazole. An efficient peptide coupling additive. *Journal of the American chemical society*, **115**, 4397–4398.

[14] Sheehan, J. C. and Hess, G. P. (1955) A new method of forming peptide bonds. *Journal of the American chemical society*, **77**, 1067–1068.

[15] Han, S.-Y. and Kim, Y.-A. (2004) Recent development of peptide coupling reagents in organic synthesis. *Tetrahedron*, **60**, 2447–2467.

[16] König, W. and Geiger, R. (1970) A new method for synthesis of peptides: activation of the carboxyl group with dicyclohexylcarbodiimide using 1-hydroxybenzotriazoles as additives. *Chemische Berichte*, **103**, 788.

[17] Angell, Y. M., García-Echeverría, C., and Rich, D. H. (1994) Comparative studies of the coupling of N-methylated, sterically hindered amino acids during solid-phase peptide synthesis. *Tetrahedron letters*, **35**, 5981–5984.

[18] Khorana, H. (1953) The chemistry of carbodiimides. *Chemical reviews*, **53**, 145–166.

[19] DeTar, D. F. and Silverstein, R. (1966) Reactions of carbodiimides I. the mechanisms of the reactions of acetic acid with dicyclohexylcarbodiimide. *Journal of the American chemical society*, **88**, 1013–1019.

[20] DeTar, D. F. and Silverstein, R. (1966) Reactions of carbodiimides II. the reactions of dicyclohexylcarbodiimide with carboxylic acids in the presence of amines and phenols. *Journal of the American chemical society*, **88**, 1020–1023.

[21] DeTar, D. F., Silverstein, R., and Rogers Jr, F. F. (1966) Reactions of carbodiimides III. the reactions of carbodiimides with peptide acids. *Journal of the American chemical society*, **88**, 1024–1030.

[22] Carraway, K. and Koshland Jr, D. (1972) Carbodiimide modification of proteins. *Methods in enzymology*, **25**, 616–623.

[23] Bauminger, S. and Wilchek, M. (1980) The use of carbodiimides in the preparation of immunizing conjugates. *Methods in enzymology*, **70**, 151–159.

[24] http://en.wikipedia.org/wiki/Carbodiimide.

[25] http://en.wikipedia.org/wiki/N-hydroxybenzotriazole.

[26] Van der Meeren, R., Wen, Y., Van Gelder, P., Tommassen, J., Devreese, B., and Savvides, S. N. (2013) New insights into the assembly of bacterial secretins, structural studies of the periplasmic domain of XcpQ from *Pseudomonas aeruginosa*. *Journal of biological chemistry*, **288**, 1214–1225.

[27] Moerman, P., Sergeant, K., Debyser, G., Devreese, B., and Samyn, B. (2010) A new chemical approach to differentiate carboxy terminal peptide fragments in cyanogen bromide digests of proteins. *Journal of proteomics*, **73**, 1454–1460.

[28] Knubovets, T., Osterhout, J. J., and Klibanov, A. M. (1999) Structure of lysozyme dissolved in neat organic solvents as assessed by NMR and CD spectroscopies. *Biotechnology and bioengineering*, **63**, 242–248.

[29] Keough, T., Lacey, M., and Youngquist, R. (2000) Derivatization procedures to facilitate *de novo* sequencing of lysine-terminated tryptic peptides using postsource decay matrix-assisted laser desorption/ionization mass spectrometry. *Rapid communications in mass spectrometry*, **14**, 2348–2356.

[30] Sergeant, K., Samyn, B., Debyser, G., and Van Beeumen, J. (2005) *De novo* sequence analysis of N-terminal sulfonated peptides after in-gel guanidination. *Proteomics*, **5**, 2369–2380.

[31] Jacob, E. and Unger, R. (2007) A tale of two tails: why are terminal residues of proteins exposed? *Bioinformatics*, **23**, e225–e230.

[32] Chung, J. J., Shikano, S., Hanyu, Y., and Li, M. (2002) Functional diversity of protein C-termini: more than zipcoding? *Trends in cell biology*, **12**, 146–150.

[33] Chung, J. J., Yang, H. M., and Li, M. (2003) Genome-wide analyses of carboxyl-terminal sequences. *Molecular & cellular proteomics*, **2**, 173–181.

Part III

Conclusion

Chapter 6

Conclusions and future perspectives

6.1 Existing methods and applications

After its important role in the determination of primary structures of proteins in early years, protein sequencing evolved in the 1980s to a method to verify the terminal regions of recombinant gene products. As presented in sections 1.8 - 1.10, this has today shifted towards large-scale structural and functional studies of the terminal regions. It has been shown that terminal protein regions have an important function in protein stability, protein traffiking, protein recognition and binding [1, 2]. These processes are often regulated by post-translational modification of the terminal regions, one of the most important being proteolytical processing [3]. The widespread shotgun proteomics approaches provide the most comprehensive identification of proteins from cellular lysates, but generally fail to characterize the N- or C-terminal sequence of the protein under study. This is due to the limited detection of terminal peptides during mass spectrometric analysis of complex peptide mixtures and the incomplete and incorrect annotation of protein termini in protein databases. As presented in Chapter 2, several terminal sequencing techniques have been developed to study protein termini. The terminal sequence information gained from these approaches has mainly been used in two relatively new 'omics' domains: so-called degradomics and proteogenomics.

Since Pehr Edman introduced his N-terminal sequencing technique in 1950, a lot of improvements and alternative techniques have been presented to characterize the terminal regions of proteins, but so far none of them has been successful enough to stop the demand for new techniques. Remarkably, most of the techniques that were discussed in the introduction have only been further exploited by the group that initially presented them. This actually means that none of them entered into a mature state for wide scale usage, or that they require specific skills and resources. Moreover, most of the techniques listed are based on a negative selection of the terminal fragment and have only been applied to rather large quantities of relatively pure protein samples. They often require multiple amino acid modifications, desalting and

purification steps to limit side reactions and the detection of false positives.

Most methods were focussed on the determination of the protein N-termini. Although less chemically reactive and hence more difficult, determining C-terminal sequences offers several advantages in proteogenomics and degradomics studies over the more popular N-terminal sequencing approaches. As many as 85% of the eukaryotic cytosolic proteins are N-terminally acetylated and therefore not easily accessible [4]. Furthermore, protein N-termini are highly susceptible to trimming by various aminopeptidases present in the sample, resulting in proteins that have N-termini missing one or more N-terminal amino acids. Due to the translation initiation mechanism, N-termini are less sequence specific, as the N-terminal amino acid is frequently a methionine residue [5]. These drawbacks are less prominent in C-terminal sequencing approaches, since protein C-termini are seldomly blocked [6] and are not as susceptible to C-terminal trimming [7].

6.2 Our contribution to the field

The starting point of this project was the C-terminal sequencing technique reported in 2005 by Bart Samyn and other members of the hosting lab for this research [8]. The aim of our work was to modify the protocol so it can be used in an automated proteome wide setup and to simultaneously improve the proteome coverage. In the existing method, carboxypeptidases were used to selectively digest the C-terminal peptide. However, the cleavage rate of these enzymes depends strongly on the amino acid sequence of the substrate: for example, Phe, Thr, Lys and Gly are cleaved off very slowly [8, 9]. To generate a peptide sequence ladder, the incubation times and CPase concentration had thus to be optimized for every sample.

In Chapter 3 we have addressed the drawbacks of the use of CPase by a novel chemical approach that allows to discriminate C-terminal peptides in CNBr mixtures. The partial ring opening of the homoserine lactone replaces the need for carboxypeptidases to identify the C-terminal peptide and makes the protocol sequence independent [10]. The method can be used for the analysis of gel or gel-free purified proteins at low femtomol sensitivity levels. By coupling the MALDI-MS/MS analysis to a robotic sample preparation device, the chemical approach proved suitable as high-throughput approach, and can be implemented in a routine proteomic setup. The technique was successfully applied to both test proteins and 96 2D-PAGE separated *Shewanella oneidensis* MR-1 proteins. The results of the 2D-PAGE experiment were compared to our previously used manual CPase ladder sequencing technique and showed a strong improvement. Moreover, we were able to identify three times more proteins using *de novo* sequenced C-terminal peptides. We have demonstrated that the technique theoretically covers 50% of the proteome of *S. oneidensis* MR-1 and that it is at least complementary to other

approaches for whole proteome C-terminal sequence determination [11]. The main limitations of the technique are intrinsic to the use of 2D-PAGE and MALDI TOF/TOF MS as analysis tools. Low ionisation efficiency of the C-terminal peptide and preferential peptide fragmentation at positively charged residues in the C-terminal peptide often reduce the quality of the MS(/MS) spectra and limit the identification rate.

Both in the CPase and chemical selection method, proteins are first cleaved with CNBr. Since methionine is a relatively rare amino acid, the peptides generated after CNBr cleavage are relatively large. In order for peptides to be efficiently detected on a MALDI-TOF/TOF analyzer they need to be in the 1-5.5 kDa mass range. Our data showed that most succesfully *de novo* interpreted MS/MS spectra had a parent ion below 3.5 kDa. This prompted us to develop a strategy to generate smaller peptides based on a method, first described by Huang and Huang, using CNBr combined with KI to cleave proteins C-terminally of Met and Trp [12]. We optimized the protocol, described the side reactions, the reaction products formed and the reaction mechanism, such that it can be used in standard proteomics experiments (Chapter 4). It proved the only protocol currently available wherein peptide bond cleavage C-terminally of Met and Trp is combined with breaking disulfide bonds in a single incubation. The need for reduction and alkylation of disulfide bridges requires to use elaborate protocols when using chemical cleavage methods, as appears from recent publications [13, 14]. Using the approach described by us, these protocols could be simplified. Because proteins are generally reduced and alkylated between the two dimensions, the application of a KI/CNBr-cleavage will offer less benefits for proteins separated by 2D-PAGE; nonetheless, the attained sequence coverage is expected to be higher than after a standard CNBr-cleavage approach. Given the interest that currently exists in increasing the sequence coverage during the study of proteins, and more specifically of membrane proteins that are often difficult to study using enzymatic digests [15–17], the chemical cleavage at two residues may be of great interest.

In an attempt to improve the proteome coverage of the existing technique, the optimized CNBr and KI cleavage protocol was implemented in the chemical selection protocol for C-terminal sequencing. We were able to distinguish the C-terminal peptides by MS and determine a C-terminal sequence tag by MS/MS [11]. Methionine accounts for 2.59% of the amino acids in *S. oneidensis* MR-1 (test species used). When CNBr is used as cleavage reagent, 51% of the C-terminal peptides can be detected and only 1 out of 3 proteins generates a C-terminal peptide in the *de novo* sequencing mass range. Tryptophan accounts for 1.25% of the amino acids in *S. oneidensis* MR-1. By cleaving C-terminal of tryptophan and methionine 58% of the C-terminal peptides are detectable on MALDI-TOF/TOF MS and 42% of the proteins have a C-terminal peptide in the *de novo* sequencing mass range. By adding KI to the reaction mixture we theoretically achieved thus a 10% larger proteome coverage (Chapter 4).

In the final results chapter, Chapter 5, we presented initial work in the development of a new MS/MS based sequencing approach. With this technique we want to overcome some of the reoccurring limitations described above. So far, only the initial reaction steps have been optimized, some suggestions have been made on possible solutions to the problems encountered.

At the start of this project the aim was to develop a proteome wide technique, able to determine a large number of C-terminal sequences. In 2007 the discussion around '2D or not 2D' was still ongoing; 2D-PAGE, not 2D-LC! At that time it seemed a logical choice to further develop the existing CPase-based C-terminal sequencing technique into a variant with higher throughput. In hindsight, the choice for a robotic platform to automate an existing technique proved not to be the best way to answer the quest for a method of fast terminomics. The technological improvements in the (LC-)MS field overwhelmed us, and we did not foresee to spend 2 to 3 years on implementing an existing technique on such an automated platform, that in the end still struggles with the same limitations and lack of sensitivity as the initial technique. Several groups, in the meantime, developed LC-MS based techniques and are able to identify several hundreds of proteins in a single, albeit complex experiment. Therefore, the PP modification based method is more likely to be successful as high throughput terminomics technique, once fully optimized.

Although our techniques might not be the most suitable for high-throughput determination of protein C-termini, they still serve a function in the proteome community. Unlike many other terminal sequencing techniques, the protocols we developed are very straightforward. They consist of very few steps that need optimalization and can be performed in less than 24 hours. As there is no need for advanced and dedicated hardware, any MS-based laboratory should be able to successfully implement our methods in their 2D-PAGE experiments. The chemical selection technique has already been applied by others to determine the C-terminal sequence of a recombinant protein [18].

6.3 Future perspectives

In most of the current sequencing techniques, proteins are cleaved into peptides and the C- or N-terminal peptides are enriched or labelled. Fragmentation of a selected peptide and *de novo* interpretation of the MS/MS spectra results in the terminal sequence of this peptide, hence of the protein. Since these cleavage reactions usually occur at the site of one or more specific amino acids, peptides of different lengths are produced. When it comes to *de novo* sequencing, mass spectrometers have a restricted optimal m/z working range, hence many of the selected peptides will not generate an interpretable MS/MS spectrum. In order to generate terminal

peptides in this optimal range, methods will need to be developed that work in combination with multiple proteases. Samples can then be analyzed in parallel generating complementary data sets.

Another way to circumvent these problems is perhaps to be found in 'top-down' proteomics, i.e. mass spectral methods to determine the sequence starting from intact proteins. Over the last few years several techniques have been reported to fragment proteins or peptides independent of their amino acid sequence in so-called non-ergodic processes [19–21]. These fragmentation methods usually generate complete series of c- and z-ions, with the post-translational modifications still intact. Since entire proteins are fragmented in top-down methods, a lot of fragment ions are generated, resulting in very complex MS(/MS) spectra. Therefore top-down methods will always require precise selection of a precurson ion representing the protein, without any interference. Currently there are no methods with enough resolution to separate proteins from a whole lysate to the necessary purity level for top-down analysis. Since proteins have very different physicochemical properties it will be very difficult to design LC-based techniques that can separate intact protein mixtures. LC-MS setups usually have a dynamic range of 10^{4th} to 10^{6th}, while *in vivo* differences in protein expression levels of 12 orders of magnitude have been observed [22]. Several affinity based techniques have been reported to deplete up to 20 of the most abundant proteins from blood serum samples, but even if this method is 99.9% successful, those 20 proteins are still 10^{9th} times more abundant than many of the important signalling proteins [23]. Not to mention that mostly a very small portion of an expressed protein is post-translationally modified, resulting in significant differences in activity.

Due to the very diverse nature of proteins, it seems very unlikely that, in the near future one technology will be developed that can successfully be applied to all protein samples, as was, for example, the case for DNA with the Sanger sequencing technology. Some techniques, like ours, focus on simple, yet sensitive and time-efficient, protocols to determine protein sequences using limited resources. Other approaches, like COFRADIC, can be applied in large scale studies and generate a large number of terminal sequences, but are time consuming and typically require a lot of instrument time. At this stage, these approaches should be regarded as complementary. Anyhow, effords to improve (high-throughput) terminal sequencing methodology are still at place. Petrera *et al.* recently pointed to the large diversity of naturally occuring carboxypeptidases. While the substrates of these peptidases are not well described, mutations and polymorphisms in the genes coding these CPases are linked to neurological and cardiovascular diseases. Terminal sequencing of proteins in mutant cell lines could help to determine their natural substrates and understand their exact pathological and fysiological function. [24].

References

[1] Sriram, S. M., Kim, B. Y., and Kwon, Y. T. (2011) The N-end rule pathway: emerging functions and molecular principles of substrate recognition. *Nature reviews molecular cell biology*, **12**, 735–747.

[2] Chung, J. J., Shikano, S., Hanyu, Y., and Li, M. (2002) Functional diversity of protein C-termini: more than zipcoding? *Trends in cell biology*, **12**, 146–150.

[3] Lange, P. F. and Overall, C. M. (2011) TopFIND, a knowledgebase linking protein termini with function. *Nature methods*, **8**, 703–704.

[4] Polevoda, B. and Sherman, F. (2000) N-α-terminal acetylation of eukaryotic proteins. *Journal of biological chemistry*, **275**, 36479–36482.

[5] Wilkins, M. R., et al. (1998) Protein identification with N and C-terminal sequence tags in proteome projects. *Journal of molecular biology*, **278**, 599–608.

[6] Nakazawa, T., Yamaguchi, M., Okamura, T.-a., Ando, E., Nishimura, O., and Tsunasawa, S. (2008) Terminal proteomics: N- and C-terminal analyses for high-fidelity identification of proteins using MS. *Proteomics*, **8**, 673–685.

[7] Xu, G., Shin, S. B. Y., and Jaffrey, S. R. (2011) Chemoenzymatic labeling of protein C-termini for positive selection of C-terminal peptides. *ACS Chemical Biology*, **6**, 1015–1020.

[8] Samyn, B., Sergeant, K., Castanheira, P., Faro, C., and Van Beeumen, J. (2005) A new method for C-terminal sequence analysis in the proteomic era. *Nature methods*, **2**, 193–200.

[9] Samyn, B., Sergeant, K., and Beeumen, J. V. (2006) A method for C-terminal sequence analysis in the proteomic era (proteins cleaved with cyanogen bromide). *Nature protocols*, **1**, 317–322.

[10] Moerman, P., Sergeant, K., Debyser, G., Devreese, B., and Samyn, B. (2010) A new chemical approach to differentiate carboxy terminal peptide fragments in cyanogen bromide digests of proteins. *Journal of proteomics*, **73**, 1454–1460.

[11] Moerman, P., Sergeant, K., Debyser, G., Timperman, I., Devreese, B., and Samyn, B. (2014) Automation of C-terminal sequence analysis of 2D-PAGE separated proteins. *EuPA open proteomics*, **3**, 250–261.

[12] Huang, S. and Huang, J. (1994) Cleavage of both tryptophanyl and methionyl peptide-bonds in proteins. *Journal of protein chemistry*, **13**, 450–451.

[13] Kuhn, K., Thompson, A., Prinz, T., Müller, J., Baumann, C., Schmidt, G., Neumann, T., and Hamon, C. (2003) Isolation of N-terminal protein sequence tags from cyanogen bromide cleaved proteins as a novel approach to investigate hydrophobic proteins. *Journal of proteome research*, **2**, 598–609.

[14] Prinz, T., Müller, J., Kuhn, K., Schäfer, J., Thompson, A., Schwarz, J., and Hamon, C. (2004) Characterization of low abundant membrane proteins using the protein sequence tag technology. *Journal of proteome research*, **3**, 1073–1081.

[15] van Montfort, B. A., Doeven, M. K., Canas, B., Veenhoff, L. M., Poolman, B., and Robillard, G. T. (2002) Combined in-gel tryptic digestion and CNBr cleavage for the generation of peptide maps of an integral membrane protein with MALDI-TOF mass spectrometry. *Biochimica et biophysica acta - Bioenergetics*, **1555**, 111–115.

[16] Laugesen, S. and Roepstorff, P. (2003) Combination of two matrices results in improved performance of MALDI MS for peptide mass mapping and protein analysis. *Journal of the american society for mass spectrometry*, **14**, 992–1002.

[17] Han, S.-Y. and Kim, Y.-A. (2004) Recent development of peptide coupling reagents in organic synthesis. *Tetrahedron*, **60**, 2447–2467.

[18] Wang, Z., Hilder, T. L., van der Drift, K., Sloan, J., and Wee, K. (2013) Structural characterization of recombinant α-1-antitrypsin expressed in a human cell line. *Analytical biochemistry*, **437**, 20 – 28.

[19] Zubarev, R. A., Kelleher, N. L., and McLafferty, F. W. (1998) Electron capture dissociation of multiply charged protein cations. A nonergodic process. *Journal of the American chemical society*, **120**, 3265–3266.

[20] Tsybin, Y. O., Witt, M., Baykut, G., Kjeldsen, F., and Hakansson, P. (2003) Combined infrared multiphoton dissociation and electron capture dissociation with a hollow electron beam in Fourier transform ion cyclotron resonance mass spectrometry. *Rapid communications in mass spectrometry*, **17**, 1759–1768.

[21] Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J., and Hunt, D. F. (2004) Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 9528–9533.

[22] Thadikkaran, L., Siegenthaler, M. A., Crettaz, D., Queloz, P. A., Schneider, P., and Tissot, J. D. (2005) Recent advances in blood-related proteomics. *Proteomics*, **5**, 3019–3034.

[23] Schuchard, M. D., Melm, C. D., Crawford, A. S., Chapman, H. A., Fan, F., Ngowe, C., Ray, K. B., Chen, D. E., and Scott, G. B. I. (2006) One step depletion of twenty high abundance human plasma proteins and concomitant molecular size fractionation of low abundance proteins. *Molecular & cellular proteomics*, **5**, S203–S203.

[24] Petrera, A., Lai, Z. W., and Schilling, O. (2014) Carboxyterminal protein processing in health and disease: key actors and emerging technologies. *Journal of proteome research*.

Curriculum vitae

Personal information

Name	Pablo Moerman
Date of birth	23/03/1985
Place of birth	Kortrijk, Belgium
Adress	Ghent, Belgium
E-mail	pablomoerman@gmail.be
Nationality	Belgian

Education

November 2007 - September 2014

PhD-candidate of Science - Biochemistry
Thesis: Novel methods for C-terminal sequence analysis in the proteome era.
Department of Biochemistry and Microbiology, Faculty of Science, Ghent University
Laboratory of protein and bimolecular engineering (L-ProBe)
Promotor: Prof. Dr. Bart Devreese
Co-Promotor: Dr. Bart Samyn
Funded by IWT-Vlaanderen

2005 - 2007

Master of Science - Biochemistry
Ghent University
Thesis: Identification of *Trichoderma reesei* proteins using charge derivatization and multidimensional liquid chromatography.
Department of Biochemistry and Microbiology, Faculty of Science, Ghent University
Laboratory of protein and bimolecular engineering (L-ProBe)
Promoter: Prof. Dr. Bart Devreese
Co-promoter: Dr. Bart Samyn

2003 - 2005

Candidate (Bachelor) of Science - Biology
Ghent University

Work

July 2013 - present

Team responsible & technical and scientific support VOC emissions
Centexbel, Zwijnaarde

Publications

Moerman P.P., Sergeant K., Debyser G., Devreese B., Samyn B. (**2010**) A new chemical approach to differentiate carboxy terminal peptide fragments in cyanogen bromide digests of proteins. *J. Proteomics* **73**, 1454-1460.

Fonseca F., Arthur C.J., Bromley E.H., Samyn B., **Moerman P.**, Saavedra M.J., Correia A., Spencer J. (**2011**) Biochemical characterization of Sfh-I, a subclass B2 metallo-beta-lactamase from *Serratia fonticola* UTAD54. *Antimicrob. Agents Chemother.* **55**, 5392-5395.

Moerman P.P., Sergeant K., Debyser G., Timperman I., Devreese B., Samyn B. (**2014**) Automation of C-terminal sequence analysis of 2D-PAGE separated proteins. *EuPA open proteomics* **3**, 250-261.

Moerman P.P., Sergeant K., Samyn B., Devreese B. (**2014**) One-step chemical cleavage of tryptophanyl and methionyl peptide bonds with concomittant oxidation of disulfide bridges, for proteomic applications. *in preparation.*

Appendices

Appendix A

Table A.1: Carboxypeptidase based C-terminal sequence analysis of 2D PAGE-separated proteins of *Shewanella oneidensis*

Protein[a]	**Spot**[b]	**Mw (kDa) /pI**	**Mw calc. (Da)**[c]	**Mw obs. (Da)**[d]	**Sequence C-term peptide**[e]
DNA-binding protein H-NS family gi|24374659	38	14.6/5.56	621.30		DDFLI
Chaperone protein DnaK gi|24372709	39	68.8/4.77	4290.05		EIAQAKAQTQGGAQEGAAKQSNATADDVVDAEFEEVKDDKK
60 kDa chaperonin GroEL gi|24372295	40,41	57.0/4.84	-		-
Enolase Eno gi|414562154	42,43	45.7/4.98	7842.12		AKAAGYTAVISHRSGETEDATIADLAVGTAAGQIKTGSLCRSDRVAKYNQLLRIEEQLGEKAPYRGLKEIKGQA
Polyribonucleotide nucleotidyltransferase PnpA gi|414561873	44	75.5/5.08	2609.39		EVDRQGRVRLSIKEAQTKEPAAE
Aconitate hydratase AcnB gi|24372027	45	93.3/5.15	9917.03		GNQARVAEGATVVSTSTRNFPNRLGTGANVYLASAELAAVAALLGRLPTVEEYQEYAKELDATAADTYRYLNFDQIDSYTKKASQVIFQSAV
Serine hydroxymethyltransferase GlyA gi|24374977	46	45.3/6.13	10451.52		LVDLIGRDLTGKEADAALGSANITVNKNSVPNDPRSPFVTSGVRIGTPAITRRGFKEAEAKQLTGWICDILDDAHNPAVIERVKGQVLALCARFPVYG
Carbamoyl-phosphate synthase small subunit CarA gi|24346790	47	41.3/5.73	8304.14		KPEALERLEKLVPVGRLGHAEEIASTVRFIIENDYVNGRVFEVDGGIRL
Glyceraldehyde-3-phosphate dehydrogenase (NAD+) GapA gi|24373892	48	36.5/5.83	331.19		NAK
Tryptophanyl-tRNA synthetase TrpS gi|24371892	49	36.7/6.19	2826.61		RAGAENAQARAEVTLKKVYEKIGLLV
Cysteine synthase CysK gi|24374432	50	34.4/5.80	6286.28		QEEGILVGISSGAAVVAANRIAALPEFADKTIVVVLPSAAERYLSSVLFQGQFGDEENIQ
Succinyl-CoA synthase α-subunit SucD gi|24373497	51	29.7/5.70	4348.38		GHAGAIIAGGKGTAADKFAALEAAGVTTVRSLADIGKALRAKTGW
Adenylate kinase Adk gi|24373581	52	23.1/5.57	10330.31		SGRRVHPGSGRVYHVVFNPPKVEGKDDVTGEDLAIRPDDEEATVRKRLGIYHEQTKPLVEYYGKVAAAGNTQYHKFDGTQSVAAVSEQLASVLK
Two component signal transduction system controlling aerobic respiration response regulator ArcA gi|24375475	53	27.2/5.51	5686.92		TGRELKPHDRTVDVTIRRIRKHFESLPDTPEIIATIHGEGYRFCGNLED
Secreted low complexity protein gi|24375332	54	34.1/5.56	1328.77		DTISGNLILKKQ
Flagellin FliC gi|24374749	55	28.5/7.90	1577.91		LAQANQLPQVALSLL
3-hydroxyacyl-CoA dehydrogenase IvdG gi|24373251	56	26.5/4.90	5502.97		KPEALERLEKLVPVGRLGHAEEIASTVRFIIENDYVNGRVFEVDGGIRL
PspF antagonist PspA gi|24373372	57	25.6/5.32	6883.58		LKFEQYERRVEGLEAQVESYDLGSKKTLADEFAALEAEDSVNAELEALKAKVKGKAPTKSKE

continued on next page

Table A.1: Carboxypeptidase based C-terminal sequence analysis of 2D PAGE-separated proteins of *Shewanella oneidensis* MR-1 – *continued from previous page*

Protein[a]	**Spot**[b]	**Mw (kDa) /pI**	**Mw calc. (Da)**[c]	**Mw obs. (Da)**[d]	**Sequence C-term peptide**[e]
TonB2 energy transduction system periplasmic component gi|414561958	58	28.9/4.94	5185.78		YNPQTKGWDKLEDSYLRELTKGIRIARKQGALDLFALPIPAAETAQ
RNA polymerase-binding protein DksA gi|24346452	59	16.8/5.13	147.08		AG
Nucleoside: proton symporter NupX gi|24375203	60	44.3/5.93	305.16		SLS

[a] Name and National Center for Biotechnology Information (NCBI)Entrez accession number.
[b] Spot number according to the position on the 2D PAGE.
[c] Mw calculated by using the residual monoisotopic values with cysteine converted to carbamidomethylcysteine.
[d] Mw observed in positive reflectron analysis (singly protonated).
[e] The sequence of the C-terminal peptide as found in the NCBI database.

Appendix B

Table B.1: Peptides resulting from the cleavage of test peptides using different halogen salts.

Salt	Sequence	Remarks
Avidin (*Gallus gallus*)		
KI	ARKCaSLTGKW	N-term. peptide
	W.KFSESTTVFTGQCaFIDRNGKEVLKTM	
	M.W°LLRSSVNDIGDDW	
	W.LLRSSVNDIGDDW	
	W.KATRVGINIFTRLRTQKE	C-term. peptide
NaI	ARKCaSLTGKW	N-term. peptide
	M.W°LLRSSVNDIGDDW	
	W.LLRSSVNDIGDDW	
	W.KATRVGINIFTRLRTQKE	C-term. peptide
2-Iodosobenzoic acid	ARKCaSLTGKW	N-term. peptide
	W.KFSESTTVFTGQCaFIDRNGKEVLKTM	
	M.W°LLRSSVNDIGDDW	
	W.LLRSSVNDIGDDW	
	W.KATRVGINIFTRLRTQKE	C-term. peptide
KCl/NaCl/4-chlorobenzoic acid	M.WLLRSSVNDIGDDWKATRVGINIFTRLRTQKE	C-term. peptide
KBr/NaBr/2-bromobenzoic acid	M.WLLRSSVNDIGDDWKATRVGINIFTRLRTQKE	C-term. peptide
α-lactalbumin (*Bos taurus*)		
KI	EQLTKCaEVFRELKDLKGYGGVSLPEW	N-term. peptide
	K.CaEVFRELKDLKGYGGVSLPEW	X-Cys(6) cleavage
	W.LAHKALCaSEKLDQW	
NaI	EQLTKCaEVFRELKDLKGYGGVSLPEW	N-term. peptide
	K.CaEVFRELKDLKGYGGVSLPEW	X-Cys(6) cleavage
	W.LAHKALCaSEKLDQW	
2-Iodosobenzoic acid	W.LAHKALCaSEKLDQW	
KCl/NaCl/4-chlorobenzoic acid	/	
KBr/NaBr/2-bromobenzoic acid	/	
Cytochrome c (*Equus caballus*)		
KI	acGDVEKGKKIFVQK	X-Cys(14) cleavage
	W.KEETLM°EYLENPKKYIPGTKM	
	W.KEETLMEYLENPKKYIPGTKM	
	M.IFAGIKKKTEREDLIAYLKKATNE	C-term. peptide

continued on next page

Table B.1: Cleavage of test peptides using different halogen salts – *continued from previous page*

Salt	**Sequence**	**Remarks**
NaI	acGDVEKGKKIFVQK	X-Cy(14)s cleavage
	W.KEETLM°EYLENPKKYIPGTKM	
	W.KEETLMEYLENPKKYIPGTKM	
	M.IFAGIKKKTEREDLIAYLKKATNE	C-term. peptide
2-Iodosobenzoic acid	acGDVEKGKKIFVQK	X-Cys(14) cleavage
	W.KEETLM°EYLENPKKYIPGTKM	
	W.KEETLMEYLENPKKYIPGTKM	
	M.IFAGIKKKTEREDLIAYLKKATNE	C-term. peptide
KCl/NaCl/4-chlorobenzoic acid	M.EYLENPKKYIPGTKM	
	M.IFAGIKKKTEREDLIAYLKKATNE	C-term. peptide
KBr/NaBr/2-bromobenzoic acid	M.EYLENPKKYIPGTKM	
	M.IFAGIKKKTEREDLIAYLKKATNE	C-term. peptide
β-lactoglobulin (*Bos taurus*)		
KI	LIVTQTM°KGLDIQKVAGTW	N-term. peptide
	M.KGLDIQKVAGTW	
	M.HIRLSFNPTQLEEQCaHI	C-term. peptide
	M.HIRLSFNPTQLEEQ	X-Cys(160) cleavage
NaI	LIVTQTM°KGLDIQKVAGTW	N-term. peptide
	M.KGLDIQKVAGTW	
	M.HIRLSFNPTQLEEQCaHI	C-term. peptide
	M.HIRLSFNPTQLEEQ	X-Cys(160) cleavage
2-Iodosobenzoic acid	LIVTQTM°KGLDIQKVAGTW	N-term. peptide
	M.KGLDIQKVAGTW	
	M.HIRLSFNPTQLEEQCaHI	C-term. peptide
KCl/NaCl/4-chlorobenzoic acid	M.KGLDIQKVAGTWYSLAM	
	M.HIRLSFNPTQLEEQCHI	C-term. peptide
KBr/NaBr/2-bromobenzoic acid	M.LIVTQTM°KGLDIQKVAGTWYSLAM	
	M.KGLDIQKVAGTW°YSLAM	N-term. peptide
	M.HIRLSFNPTQLEEQCHI	C-term. peptide

All terminal Met or Trp residues were observed in the modified homoserine lactone or Cγ-O-spirolactone form.
°= Oxidation of Met of Trp.
ac = acetylation
Ca = Cysteic acid.